DOCUMENT RESUME

ED 171 730 TM 008 622

AUTHOR Warm, Thomas A.
TITLE A Primer of Item Response Theory. Technical Report 940279.
INSTITUTION Coast Guard Inst., Oklahoma City, Okla.
REPORT NO CG-941278
PUB DATE Dec 78
NOTE 155p.
AVAILABLE FROM National Technical Information Service, Springfield, Virginia 22161 (AD-A063072, MF $3.00, HC $8.00)

EDRS PRICE MF01/PC07 Plus Postage.
DESCRIPTORS *Ability Identification; Achievement Tests; Aptitude Tests; Bayesian Statistics; Computer Programs; Cutting Scores; Equated Scores; *Item Analysis; Multiple Choice Tests; Probability; *Psychometrics; *Scoring Formulas; *Statistical Analysis; Test Bias; *Test Interpretation
IDENTIFIERS Adaptive Testing; Item Characteristic Curve Theory; *Item Response Theory; Latent Trait Theory; Maximum Likelihood Estimation; Tailored Testing

ABSTRACT

This primer is an introduction to item response theory (also called item characteristic curve theory, or latent trait theory) as it is used most commonly--for scoring multiple choice achievement or aptitude tests. Written for the testing practitioner with minimum training in statistics and psychometrics, it presents and illustrates the basic mathematical concepts needed to understand the theory. Then, building upon those concepts, it develops the basic concepts of item response theory: item parameters, item response function, test characteristic curve, item information functions, test information curve, relative efficiency curve, and score information curve. The maximum likelihood and Bayesian modal estimates of ability are described with illustrative examples. After a discussion of assumptions and available computer programs, some practical applications are presented, i.e. equating scales, tailored testing, item cultural bias, and setting pass-fail cut-offs. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED171730
IM008 622

DEPARTMENT OF TRANSPORTATION

COAST GUARD

TECHNICAL REPORT
940279

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

COAST GUARD INSTITUTE
OKLAHOMA CITY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Thomas A. Warm

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A PRIMER OF ITEM RESPONSE THEORY

Thomas A. Warm

U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, Oklahoma 73169

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
CG-941278	AD-AØ63Ø72	

4. Title and Subtitle	5. Report Date
A Primer of Item Response Theory	DECEMBER 1978
	6. Performing Organization Code
	8. Performing Organization Report No.
7. Author(s) T.A. Warm	
9. Performing Organization Name and Address U.S. Coast Guard Institute P. O. Substation 18 Oklahoma City, OK 73169	10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered 14. Sponsoring Agency Code

15. Supplementary Notes
Copies of this book may be obtained from the National Technical Information Service or the Defense Documentation Center by sending $8.oo for papercopy or $3.oo for microfische. Use item #AD-AØ63Ø72 on order form on page 2.

16. Abstract

This book is an introduction to Item Response Theory (IRT) (also called Item Characteristic Curve Theory, or latent trait theory).

It is written for the testing practitioner with minimum training in statistics and psychometrics. It presents in simple language and with examples the basic mathematical concepts needed to understand the theory.

Then, building upon those concepts, it develops the basic concepts of Item Response Theory: item parameters, item response function, test characteristic curve, item information functions, test information curve, relative efficiency curve, and score information curve. The maximum likelihood and Bayesian modal estimates of ability are described with illustrative examples. After a discussion of assumptions and available computer programs, some practical applications are presented, i.e. equating scales, tailored testing, item cultural bias, and setting pass-fail cut-offs.

17. Key Words	18. Distribution Statement
item response theory, item characteristic curve theory, latent trait theory, item analysis, test theory, cultural bias, tailored testing, adaptive testing	Distribution unlimited.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	159	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

MAIL ORDER TO:

NTIS

National Technical Information Service
U.S. DEPARTMENT OF COMMERCE
Springfield, Va. 22161
(703) 557–4650 TELEX 89–9405

For DDC Users Only

DDC User Code _____

Contract Number _____
(last 6 characters only)

Treasury Agency Location Code
For Government Users
(who report on SF–224)
(8 digit)

☐☐ ☐☐ ☐☐☐☐

Date _____

PURCHASER: Telephone: _____

SHIP TO:
(Enter if different from address at left)

Name _____

Organization _____

Address _____

City, State, ZIP _____

Attention: _____

METHOD OF PAYMENT

- ☐ Charge my NTIS deposit account no. _____
- ☐ Purchase order no. _____
- ☐ Check enclosed for $ _____
- ☐ Ship & Bill. See reverse (not applicable outside North America).
- ☐ Charge to my American Express Card account number $12.50

Card expiration date _____

Signature _____

Order Options

Normal delivery time takes three to five weeks. It is vital that you order by number or your order will be manually filled, insuring a delay. You can opt for *airmail delivery* for a $3.00 charge per item. Just check the *Airmail Service* box. If you're really pressed for time, call the NTIS Rush Handling Service (703) 557–4700. For a $10.00 charge per item your order will be airmailed within 24 hours. Or, you can pick up your order in the Washington Information Center & Bookstore or at our Springfield Operations Center within 24 hours for a $6.00 per item charge.

ITEM NUMBER ***	USER ROUTING CODE (see reverse)	Quantity: Paper Copy	Quantity: Microfiche	Quantity: Other (specify)	UNIT** PRICE	AIRMAIL* SERVICES	TOTAL PRICE
AD-A063072							

* Add $2.00 per item for First Class Delivery in North America;
Add $3.00 for each paper copy Airmail Delivery outside North America;
Add $.75 for each microfiche Airmail Delivery outside North America.
** ALL PRICES SUBJECT TO CHANGE.
*** If ordering by title or if item ordered is a magnetic tape, please see reverse side.

Enter Grand Total $

5

NTIS–173 (4/77)

TABLE OF CONTENTS

Page	Chap	
3		Table of Contents
5		Glossary of Special Terms and Symbols
7		Bookmark and Glossary
9		Preface
11	1	Introduction
15	2	Classical Test Theory vs. Item Response Theory
19	3	A Brief History of Item Response Theory
23	4	The Normal Ogive and Logistic Ogive
27	5	More About Logistic Ogives
39	6	The Item Response Function (IRF)
51	7	The a, b, & c Parameters
57	8	The Test Characteristic Curve
63	9	The Item Information Function (IIF)
72	10	The Test Information Curve and Relative Efficiency Curve
79	11	The Score Information Curve
81	12	Maximum Likelihood Estimation of Θ
93	13	Bayesian Modal Estimation of Θ
97	14	Assumptions
109	15	Computer Programs
113	16	Equating the Θ Scales
119	17	Tailored Testing
125	18	Item Cultural Bias
143	19	Setting Minimum Passing (Cut-Off) Scores
149		Postword
151	App. A	Logistic Identities & Equations
153		References

Note: The following pages are intentionally blank: 4, 6, 14, 18, 33, 50, 71, 78, 85, 92, 108, 142, 150, 152.

Glossary of Special Terms and Symbols

	= # of alternatives in a multiple choice question
a-value	= discrimination index
ASI	= Alternative Similarity Index
b-value	= difficulty index
BME	= Bayesian Modal Estimation
c-value	= pseudo-guessing index
CRT	= Cathode Ray Tube device
d-value	= point biserial correlation
d.f.	= distribution function, an ogive
E	= Error score
e	= base of natural logarithm
exp()	= e raised to the power of whatever is in the parenthesis after the exp
f.f.	= frequency function, bell shaped curve
I(θ)	= Test Information Curve
I(θ,u)	= Item Information Function
ICC	= Item Characteristic Curve, same as IRF
IIF	= Item Information Function, I(θ,u)
IRF	= Item Response Function
IRT	= Item Response Theory
KR-20	= Kuder-Richardson Formula 20
	= Likelihood
L(0,1.7)	= Logistic Frequency Function
L(θ|U)	= Likelihood of θ, given U
L(U|θ)	= Likelihood of U, given θ
m	= slope of the ogive at the b-value
MAPL	= Minimum Acceptable Performance Level
MLE	= Maximum Likelihood Estimation
N(0,1)	= Normal f.f.
p-value	= proportion of examinees selecting an item alternative
P_i	= $P_i(\theta)$ = Probability of getting item correct, given θ
Q_i	= $Q_i(\theta)$ = Probability of getting item wrong, given θ
$r_{g\theta}$	= item biserial correlation
r_{gh}	= interitem tetrachoric correlation
r_{xx}	= reliability of classical test theory
REC	= Relative Efficiency Curve, ratio of TIC's
SBAYES	= Simplified Bayesian, same as BME
SD	= standard deviation
SEE	= Standard Error of Estimate
SIC	= Score Information Curve
SME	= Subject Matter Expert
T	= True score, Observed score - Error
TIC	= Test Information Curve, $I(\theta)$, $\Sigma I(\theta,u)$
USCSC	= U.S. Civil Service Commission
U	= response vector, response pattern
u	= response, $u_i = 1$ if response is correct & $u_i = 0$ if response is wrong
W(θ)	= optimal weight of an item
X	= Observed score
$\bar{X}$	= Mean
θ	= Theta, the ability scale
$\int$	= Integral sign
Ψ	= Psi, logistic ogive
Φ	= Phi, normal ogive
Σ	= Summation of a series of numbers
Π	= Product of a series of numbers

Fold and Tear Here

BOOKMARK AND GLOSSARY

of special terms and symbols

- A = # of alternatives in a multiple choice question
- a-value = discrimination index
- ASI = Alternative Similarity Index
- b-value = difficulty index
- BME = Bayesian Modal Estimation
- c-value = pseudo-guessing index
- CRT = Cathode Ray Tube device
- d-value = point biserial correlation
- d.f. = distribution function, an ogive
- E = Error score
- e = base of natural logarithm
- exp() = e raised to the power of whatever is in the parenthesis after the exp
- f.f. = frequency function, bell shaped curve
- $I(\theta)$ = Test Information Curve
- $I(\theta,u)$ = Item Information Function
- ICC = Item Characteristic Curve, same as IRF
- IIF = Item Information Function, $I(\theta,u)$
- IRF = Item Response Function
- IRT = Item Response Theory
- KR-20 = Kuder-Richardson Formula 20
- L = Likelihood
- L(0,1.7) = Logistic Frequency Function
- $L(\theta|U)$ = Likelihood of θ, given U
- $L(U|\theta)$ = Likelihood of U, given θ
- m = slope of the ogive at the b-value
- MAPL = Minimum Acceptable Performance Level
- MLE = Maximum Likelihood Estimation
- N(0,1) = Normal f.f.
- p-value = proportion of examinees selecting an item alternative
- P_i = $P_i(\theta)$ = Probability of getting item correct, given θ
- Q_i = $Q_i(\theta)$ = Probability of getting item wrong, given θ
- $r_{g\theta}$ = item biserial correlation
- r_{gh} = interitem tetrachoric correlation
- r_{xx} = reliability of classical test theory
- REC = Relative Efficiency Curve, ratio of TIC's

over

SBAYES = Simplified Bayesian, same as BME
SD = standard deviation
SEE = Standard Error of Estimate
SIC = Score Information Curve
SME = Subject Matter Expert
T = True score, Observed score - Error
TIC = Test Information Curve, $I(\theta)$, $\Sigma I(\theta,u)$
USCSC = U.S. Civil Service Commission
U = response vector, response pattern
u = response, $u_i = 1$ if response is correct & $u_i = 0$ if response is wrong
$W(\theta)$ = optimal weight of an item
X = Observed score
$\bar{X}$ = Mean
θ = Theta, the ability scale
$\int$ = Integral sign
Ψ = Psi, logistic ogive
Φ = Phi, normal ogive
Σ = Summation of a series of numbers
π = Product of a series of numbers

PREFACE

One year ago I had never heard of latent trait theory, an item characteristic curve, or Fred Lord. On my first reading of Lord and Novick (1968) Chapters 16 and 17, I understood absolutely nothing. After several hours of study on my second reading, I finally comprehended one simple equation. During the next several months I reread parts of Lord and Novick as many as 20 times, I taught myself some differential calculus, integral calculus, mathematical statistics, probability theory and linear algebra, I attended Fred Lord's course in Item Response Theory at the Educational Testing Service, Princeton, NJ, and I read several publications on Item Response Theory.

I have now gotten to the point where I am able to use Item Response Theory for my purposes, although there is still much that I do not understand.

Upon reflection, I find that, as is true in many sciences, it is not necessary to fully understand the theoretical background and mathematical development in order to apply the results of the model.

It is widely acknowledged in the field that one of the main reasons that item response theory has been so slow to catch on among testing practitioners is the mathematical complexity of the literature. Most of the literature is written with language and notation that is standard for the researchers. However, that language and notation is confusing to the thousands of testing practitioners, whose technical training amounts to a couple of courses in statistics and tests and measurement, if that much. On the other hand, many of the concepts used in the literature are not difficult to understand, if explained in less esoteric language and with a few examples.

Therefore, it became my resolve that no testing practitioner, such as I, should have to go through what I went through in order to gain a basic understanding of item response theory. The purpose of this paper is to fulfill that resolve.

Since very little of this paper is original with me, by rights there should be a reference for nearly every sentence or paragraph. Such complete references, however, will not be included because they would be out of place for a primer, and usually not of interest to the novice. My primary references are Lord & Novick (1968) and Lord (in preparation). Some references will be included to direct the reader to more thorough and detailed explanations. Other references will be included where authoritative support is deemed desirable.

A primer is necessarily incomplete. It is also inaccurate when it contains oversimplifications which apply to the general case, but do not apply to extreme, unusual, or uninteresting cases. This paper will be guilty of such generalities and rules of thumb.

Other excellent, less elementary introductory material is also available. (See Baker, 1977; Hambleton & Cook, 1977; Sympson, 1977).

I am indebted to ENS Debra Cook, ENS Pamela Crandall, ENS Charles Pastine, and LTJG Larry Young for their assistance in the analysis of data.

My appreciation for the many suggestions and corrections made by the several readers and reviewers is gratefully acknowledged. They are: John A. Burt, Joseph Cowan, Myron A. Fischl, Steven Gorman, Karen Jones, Frederick M. Lord, James R. McBride, W. Alan Nicewander, Malcolm J. Ree, and James B. Sympson.

I would also like to thank YN2 Ron Smith for his excellent art work, and Jim Walls for his systems analysis and computer programming.

THOMAS A. WARM

January 22, 1978

CHAPTER 1
INTRODUCTION

1.1 Item Response Theory (IRT) is the most significant development in psychometrics in many years. It is, perhaps, to psychometrics what Einstein's relativity theory is to physics. I do not doubt that during the next decade it will sweep the field of psychometrics. It has been said that IRT allows one to answer any question about an item (test question), a test, or an examinee, that one is entitled to ask. Although this statement is somewhat circular, it will give you an idea of the terrific power of IRT and of the mathematical estimation methods involved.

The most common application of IRT is with multiple-choice questions in an ability test. That use will be the thrust of this paper, although IRT also applies as well to free response (fill in) items. I make no distinction between ability and knowledge testing. IRT applies to tests for both. Thus, the word "ability" will be used for both types of tests. No application of IRT to personality or interest testing will be discussed.

1.2 If we give several tests in the same subject matter area to a group of examinees, we find that in general the same examinees score high on the tests and the same examinees score low. In other words, we find consistency in the performance of examinees on the different tests.

To explain this consistency we assume that there is something inside the examinees that causes them to score consistently. We call that something a mental trait.

In the vernacular the word "trait" implies an innate, inherited characteristic. We don't necessarily mean that. We mean only that characteristic of the examinee that causes consistent performance on the tests, whatever, if anything, it is.

No one has found a physical referent for a mental trait, and few really expect to. It is sometimes tempting to think of a trait as having a physical referent like a brain engram, but that is always unnecessary. In this sense, a trait is an intervening variable, as opposed to a hypothetical construct. Since the mental trait has no known physical referent, it is never observed directly. Therefore, it is called a "latent" trait.

1.3 The scale of the latent trait is traditionally given the name of the Greek letter theta (θ). I will use the terms theta, ability level, amount of trait, and amount of subject-matter-knowledge, interchangeably. Theta is a continuum from minus infinity ($-\infty$) to plus infinity ($+\infty$). It has no natural zero point or unit. Therefore, the zero point and unit are often taken as the mean and standard deviation, respectively, of some reference sample of examinees. Thus, values of θ usually vary from -3 to +3, but may be observed outside that range. The θs of a sample need not be distributed normally.

1.4 When an examinee walks into a testing room, he brings with him his theta.* The purpose of the test, then, is to measure the relative position of the examinees on the theta scale. The test interprets the examinee's theta and produces a measurement of ability, which is often the raw (number right) score. The test is the measuring instrument. Often measurement of an ability with a test is made analogous to measurement of height with a tape rule. But there is an important difference. Height, whether measured by an English rule or metric rule, is always on an equal interval scale. Histograms of a group of people will always look the same, except for some linear stretching of a scale.

*The generic masculine pronouns will be used for convenience.

That is not the case with testing. The histograms of raw scores of the same people on two tests will seldom look the same, even with linear stretching of a scale. That is because each test has its own peculiar scale (also called metric). The peculiarity of a test's metric distorts the distribution of examinees. Until IRT there has been no way to identify the peculiar scale of a test.

CHAPTER 2

Classical Test Theory vs. Item Response Theory

2.1 Classical test theory has been developed over a period of many years. Gulliksen (1950) is an excellent presentation of classical test theory.

Most testing practitioners use classical test theory, whether they know it or not. The basic tools of most testing practitioners are:

a. p-value = proportion of examinees selecting an item alternative (also called "item difficulty"),

b. d-value = point-biserial correlation between the item alternative and the test (some use the biserial correlation)(also called "item discrimination"),

c. mean of examinees' (number right) scores,

d. standard deviation of examinees' scores,

e. skewness and kurtosis of examinees' scores,

f. reliability of the test, usually KR-20, the Kuder-Richardson Formula 20 (a special case of Cronbach's coefficient alpha).

Anyone whose test analysis is principally based on the statistics listed above is using classical test theory. The problem with those statistics is that they are relative to the characteristics of the test and of the examinees.

The p-value is relative to the ability level of the examinees. The same item given to a high ability group and low ability group will get two different p-values for the two groups. It can be shown that p-values are not true measures of relative item difficulty. It is not uncommon for items measuring the same ability to reverse the order of their p-values when given to groups of different average ability. For example, item A may have a higher p-value than item B for one group of examinees, but have a lower p-value than item B for a different group. This effect is not a matter of sampling error.

The d-value is relative to the homogeneity of the ability levels of the examinees in the sample, the subject-matter homogeneity of the items in the test, and the dispersion of p-values of items in the test. The same item, given to a group of examinees who are similar in ability and to another group with a wide range of ability, will produce two different d-values for the two groups. Similarly, an item included in a test with other items that are homogeneous in content and p-value will get a d-value different from the d-value it will receive in a heterogeneous test.

The mean, standard deviation, skewness and kurtosis will also vary according to the characteristics of the test and examinees.

The reliability is relative to the standard deviation of the test, and to the p-values and d-values of the items in the test, all of which are dependent upon the particular abilities of the examinees and the characteristics of the test.

The following quote gives another liability of using classical test theory in culture-fair testing studies:

"It can be shown that classical parameters (e.g. p-value) will generally not be linearly related across subgroups of a population. This means that the test for cultural bias using classical parameters can lead to an artifactual detection of bias." (Pine, 1977, p.40)

Clearly, classical test theory statistics are meaningful only in an extremely limited situation, i.e., when the same item is given to the same population as part of strictly parallel tests. Such a situation rarely occurs. Furthermore, the basic precepts and definitions of classical test theory are untestable, i.e. they are tautologies. They are simply taken as true without any way to empirically determine their relevance to reality. Some are assumed to be true even when this does not appear to be warranted. Thus, no one knows if the classical test model applies to any real test.

2.2 In contrast IRT makes possible item and test statistics which are dependent neither on the characteristics of the examinees nor on the other items in the test. They are invariant. With the item statistics it becomes possible to describe in precise terms the characteristics of the test before the test is administered. This capability allows one to construct a test that is highly efficient in accomplishing the purpose of the test. It also provides an extremely powerful tool for special studies, such as item cultural bias.

Moreover, the assumptions of IRT are explicit and have the potential of empirical testing. It is possible to discover if the data reasonably meet the assumptions.

CHAPTER 3

A Brief History of Item Response Theory

3.1 The origin of latent trait theory can be traced to Ferguson (1942) and Lawley (1943). Item Response Theory is just one of several models under latent trait theory. The Rasch model is another.

3.2 Other early publications using some of the same concepts are Brogden (1946), Tucker (1946) Carroll (1950), and Cronbach and Warrington (1952).

3.3 In 1952, Lord published his Ph.D. dissertation in which he presented IRT as a model or theory in its own right. At that time he called it Item Characteristic Curve Theory. Thus, Lord is considered the father and founder of IRT. Shortly after publishing his dissertation, Lord stopped work on IRT for ten years, due to a seemingly intractable problem with it.*

3.4 In 1960, Rasch (1960) published his one-parameter sample-free model. The Rasch model stirred much interest and considerable work was done on it during the next decade. Its leading proponent in the U.S. is Benjamin Wright, a psychoanalyst at the University of Chicago. (See Wright, 1977 for references).

3.5 In 1965, Lord (1965) conducted a massive study, using a sample size of greater than 100,000. That study showed that the "problem", which had deterred his work for so long, was not really a problem, and that IRT was appropriate for real life multiple-choice tests. With that study Lord began work again on IRT.

*This problem is discussed in Section 14.2

3.6 In 1968, Lord and Novick published a psychometrics textbook, within which were four Chapters (17-20) by Allan Birnbaum (1968), a well-known statistician (now deceased). Birnbaum's chapters worked out in detail the mathematics of the two and three parameter normal ogive and logistic models.*

3.7 Soon thereafter Urry (1970) completed his Ph.D dissertation in which he compared the one, two, and three parameter models. He concluded that the three parameter model best described the real world for multiple-choice tests.

3.8 Since Urry's dissertation, much work has been done on all three models (i.e., one, two, and three parameter), but the three parameter model is now receiving most of the attention because it best describes reality. To wit, I shall deal with the 3-parameter model only.

3.9 Much of the work on the 3-parameter model is coming from 3 principal sources. The sources are:

a. Frederic M. Lord, Distinguished Research Scientist, Educational Testing Service, Princeton, NJ.

b. Vern W. Urry, Personnel Research Psychologist, United States Civil Service Commission, Washington, D.C.

c. David J. Weiss, Prof. of Psychology, Psychometric Methods Program, University of Minnesota, Minneapolis, MN.

There are, of course, many other highly productive researchers publishing excellent studies. Failure to include them in this list is more an indication of my limited exposure than of the significance of their contributions.

*The normal ogive and logistic ogive will be compared briefly in Chapter 4.

3.10 The United States Civil Service Commission has adopted a particular application of IRT as official policy. The five U.S. armed forces (including the U. S. Coast Guard) are also investigating the application of IRT.

3.11 In 1977 Lord changed the name of his model from Item Characteristic Curve Theory to Item Response Theory.

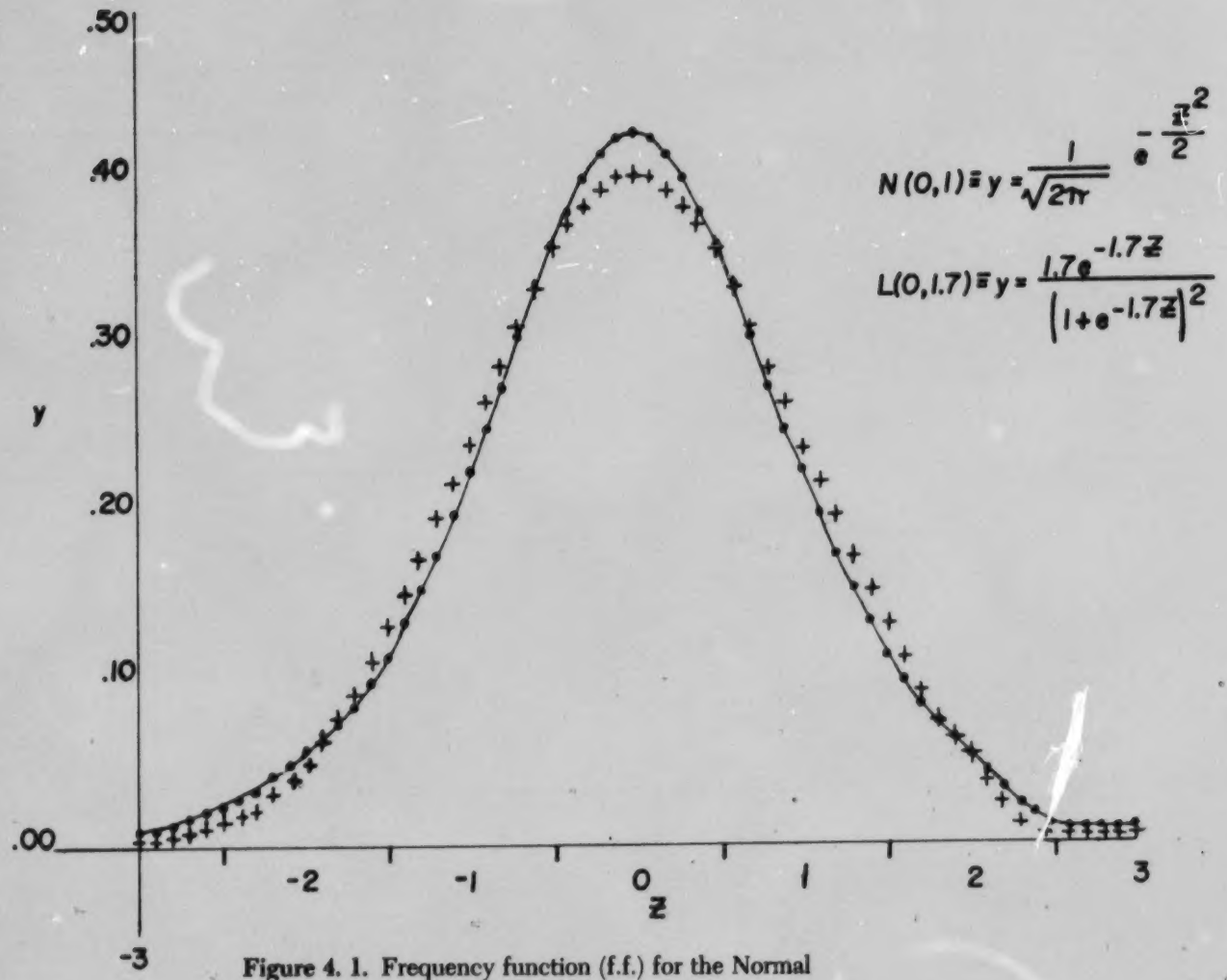


Figure 4. 1. Frequency function (f.f.) for the Normal Curve (+ + + +) and Logistic Curve (....).

CHAPTER 4

The Normal Ogive and Logistic Ogive

4.1 I trust the reader will recognize the normal curve plotted in Figure 4.1 with the pluses (++++). It has a mean =0, and standard deviation =1. The formula for this normal curve is identified in Figure 4.1 as N(0,1).

4.2 A bell-shaped curve like this is called a frequency function (f.f.). It is called a frequency function even when the ordinate (vertical axis) is defined as frequency, proportion, percent, or density (Kendall and Stuart, 1977, p. 13). Therefore, we call the normal curve, the "normal frequency function."

4.3 Superimposed over the normal f.f. in Figure 4.1 is a logistic* curve or logistic frequency function, plotted with dots (.....). This logistic f.f. also has a mean =0 and standard deviation ≈ 1.0. The formula for this logistic f.f. is identified in Figure 4.1 as L(0,1.7). The 1.7 in the exponent of the formula is chosen to allow the logistic f.f. to approximate the normal f.f as closely as possible. The actual value is 1.6679, which is rounded to 1.7. In some of the literature the 1.7 is represented by the upper case letter D. The letter e is the base of natural logarithms; $e \approx 2.718281828$.

4.4 The reader will also recognize the S-shaped curve in Figure 4.4 as the normal cumulative frequency curve. An S-shaped curve is called an ogive.** This curve gives the proportion of area under the normal curve (Figure 4.1) that lies to the left of each point on the abscissa (horizontal axis).

*pronounced lojistic
**pronounced ojive

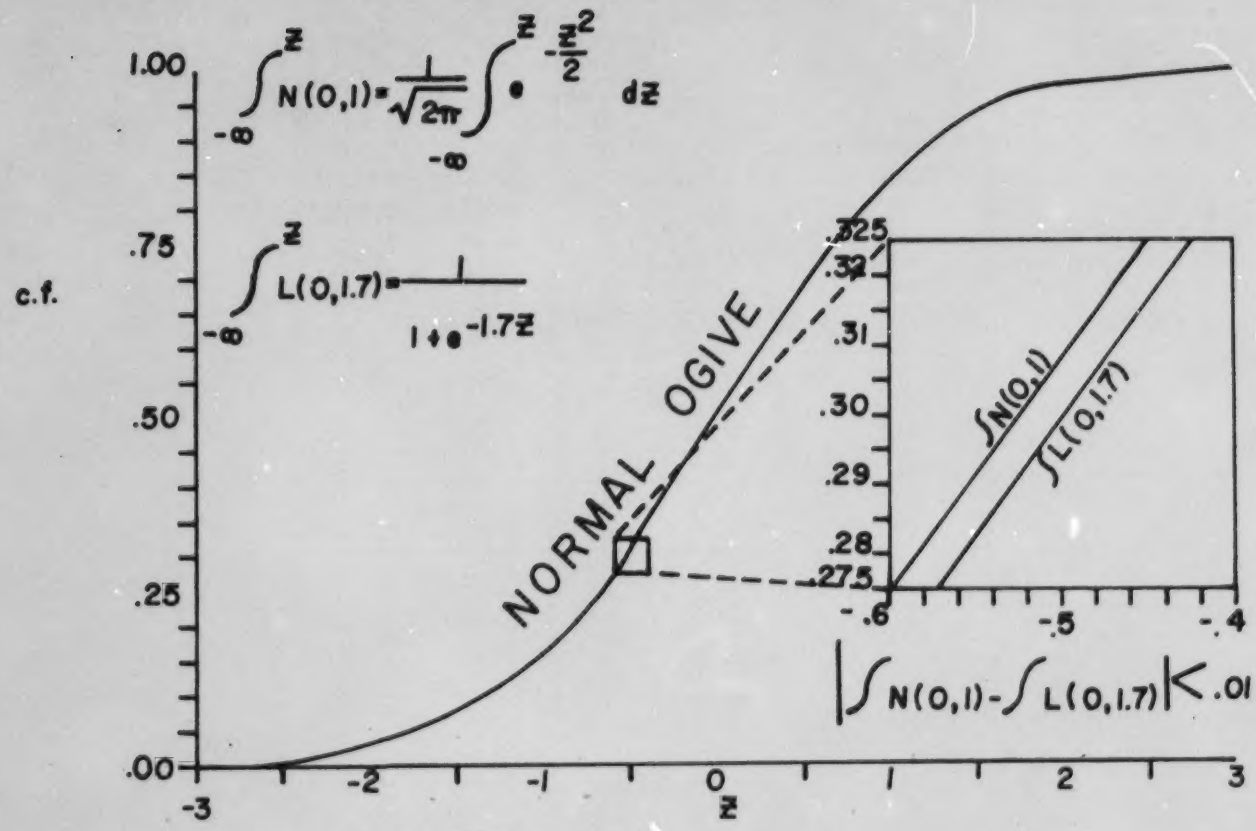


Figure 4.4. Distribution function (d.f.)for the N (0,1) and L (0,1.7) frequency functions.

4.5 An ogive like this is called a distribution function (d.f.). It is called a distribution function even when the ordinate is defined as cumulative frequency, cumulative proportion, cumulative percent, or cumulative area (Kendall & Stuart, 1977, p.13). Therefore, we call the curve in Figure 4.4 a "normal distribution function," or a "normal ogive". The formula for this normal d.f. is identified in Figure 4.4 as $\int N(0,1)$.

4.6 Also in Figure 4.4, but not discernable, is the logistic ogive (or logistic d.f.) for the logistic f.f. in Figure 4.1. It is not discernable, because it is so close to the normal ogive that on this scale the two curves merge together in the width of the ink line. A small portion has been magnified to a larger scale (10x), so that the difference may be seen. The magnified area was chosen at the place where the 2 ogives are farthest apart. The reader can verify that at any point on the abscissa the 2 ogives are always less than .01 apart on the ordinate, as is indicated by the inequality under the magnification in Figure 4.4. The formula for this logistic d.f. is identified in Figure 4.4 as $\int L(0,1.7)$.

4.7 The ogive with which we are concerned is the normal ogive. However, note the integral sign ($\int$) on the right side of the definition for the $\int N(0,1)$.

The integral sign there means that no algebraic function can be found to describe the normal ogive. This fact makes the normal ogive very cumbersome to work with mathematically, and requires numerical methods to solve, or a table of values.

4.8 On the other hand the logistic ogive has no integral sign on the right side of its definition ($\int$ L(0,1.7)). In fact, the expression on the right in Figure 4.4 is the algebraic function describing the logistic ogive. The logistic ogive is very easy to work with.*

4.9 For these reasons the logistic ogive is substituted as a convenient and very close approximation to the normal ogive.

4.10 This paper will only deal with the logistic ogive. Statements about the logistic ogive may be taken as close approximations to the normal ogive model. The logistic f.f. is no longer of interest to us.

*Some interesting logistic identities are given in Appendix A.

CHAPTER 5
More About Logistic Ogives

5.1 Figure 4.4 shows just one logistic ogive. There is actually an infinite family of logistic (and normal) ogives, each different in some way from every other one.

5.2 Logistic ogives are strictly monotonic functions. They are strictly monotonic because, going from left to right, the ogive always gets higher and higher, never is completely horizontal, and never goes down.

5.3 Notice the ogive in Figure 4.4. Between -2.0 and -0.5 on the horizontal axis the ogive is concave upward. Between 0.5 and 2.0 it is concave downward. At some point between -0.5 and 0.5 this ogive must change from being concave upward to concave downward. That point is called the "inflection point." The inflection point is always the point where the slope of the ogive is at its maximum. The inflection point for this ogive is located on the vertical axis at .50, and on the horizontal axis at 0.0.

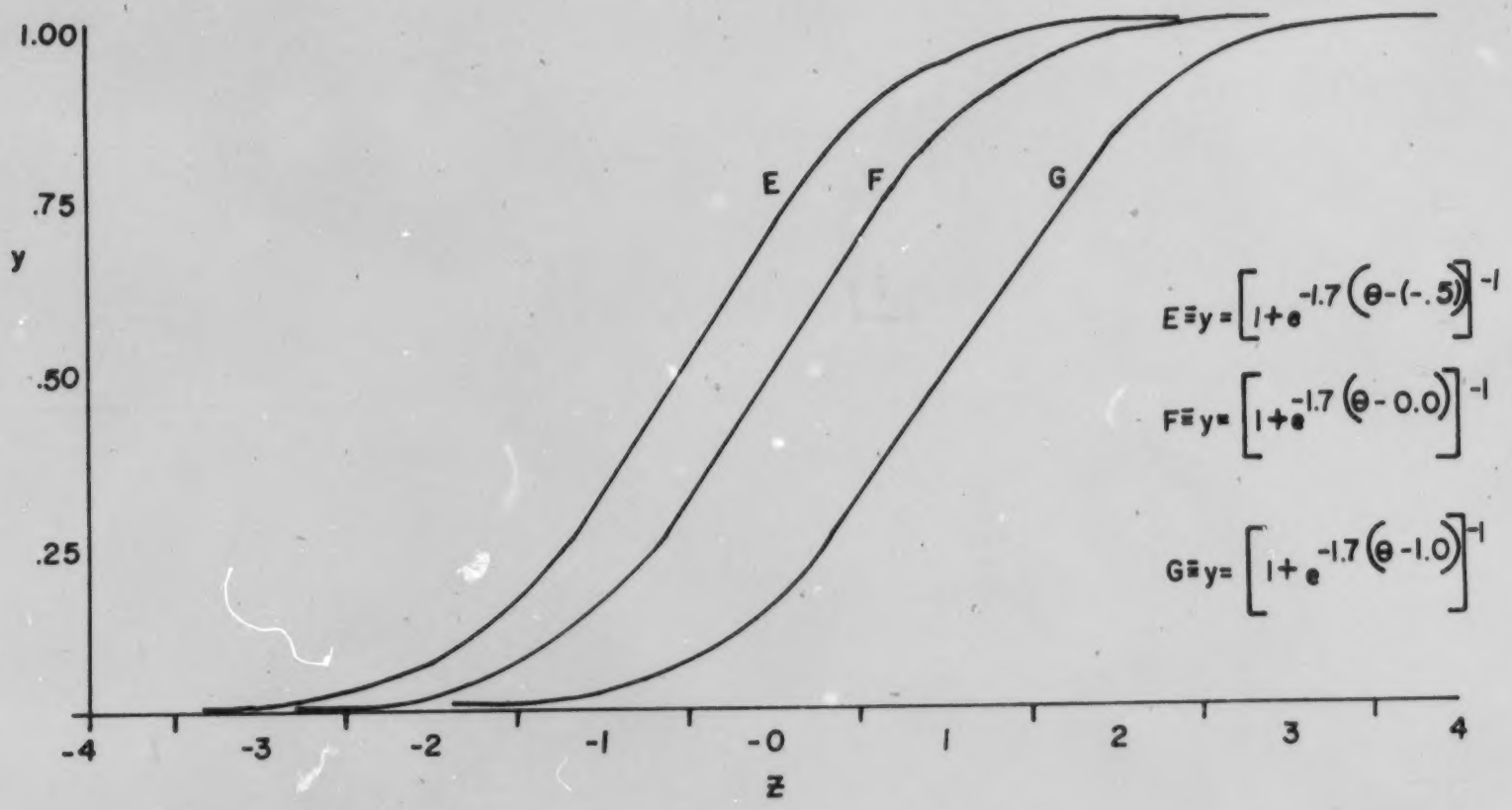


Figure 5.5. Three logistic ogives (E,F, and G) with b = -.5, 0.0, and 1.0 respectively.

5.4 Three-parameter logistic ogives (with which we are exclusively concerned) may differ from each other in only 3 ways, one for each parameter.

5.5 One way in which logistic ogives may differ is in the horizontal location of the inflection point. Figure 5.5 shows 3 logistic ogives labeled E, F, and G with their inflection points at different places on the abscissa. You can see that the 3 ogives are exactly the same except for a sideways shift of the entire curve. Shifting the inflection point sideways, shifts the entire ogive sideways. The horizontal position of the inflection point is called the "b-parameter". Some call it, as we will, the "b-value". The b-values of ogives E, F, and G in Figure 5.5 are -.5,0.0 and 1.0, respectively.

5.6 To include the b-parameter in the logistic ogive function, it is only necessary to subtract the b-parameter from the horizontal axis variable.

5.7 Figures 4.1, 4.4, and 5.5 were constructed with the horizontal axis labeled z. This label was chosen to facilitate understanding of the logistic f.f and d.f., because of the reader's likely familiarity with the traditional z-scores of measurement. Since we are concerned with the ability scale called θ, we now and hereafter label the horizontal axis, θ. Substituting θ for z in the logistic function and subtracting the b parameter, gives the height of the logistic ogive by the function

$$\Psi(\theta) = \left[1 + e^{-1.7(\theta - b)}\right]^{-1}$$

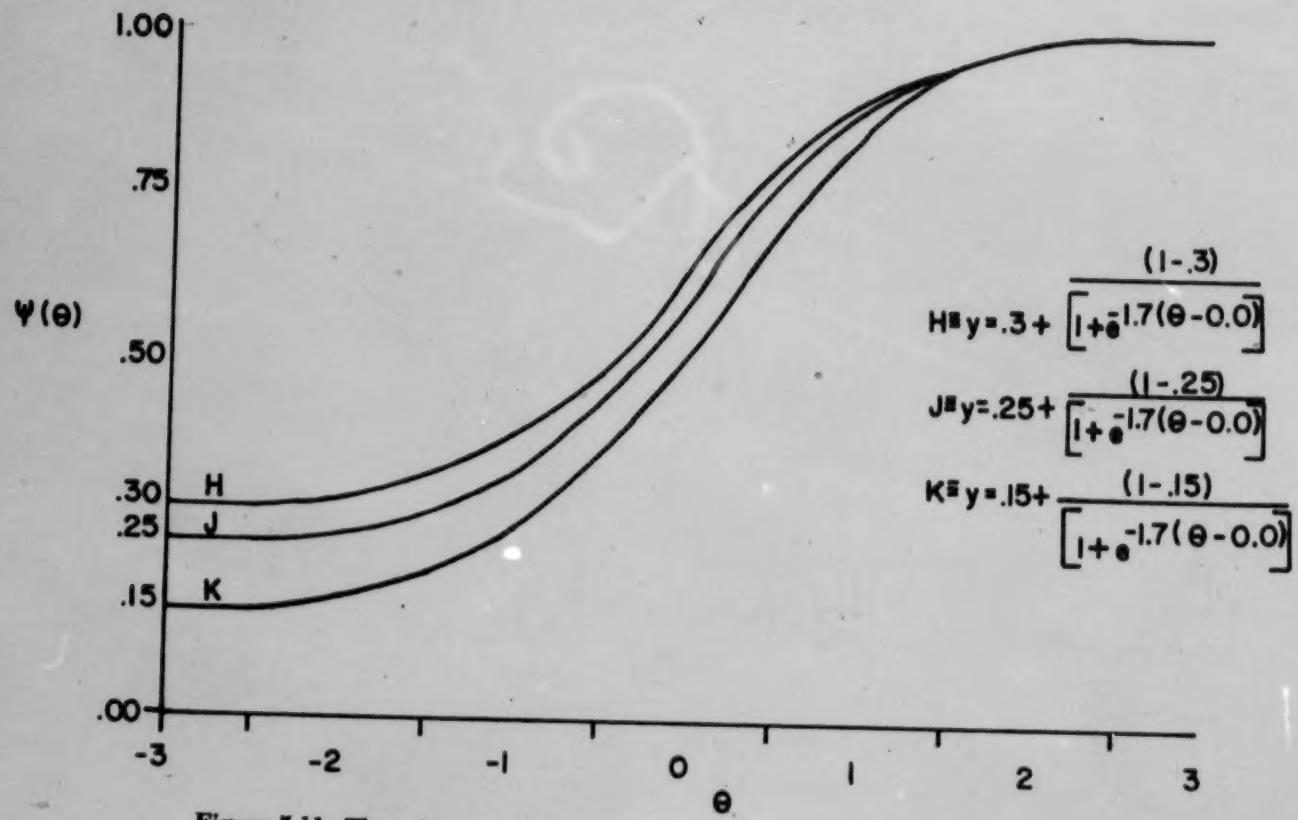


Figure 5.11. Three logistic ogives (H, J, and K) with b = 0.0 and c = .30, .25, and .15 respectively.

which is sometimes written

$$\Psi(\theta) = \left[1+\exp(-1.7(\theta-b))\right]^{-1}$$

where exp means e raised to the power of whatever is in the parenthesis after the exp. The upper case Greek letter psi (Ψ) is used in the literature to mean the logistic ogive. Phi (Φ) is used to mean the normal ogive.

5.8 The logistic ogive has 2 asymptotes. The asymptotes are horizontal lines that the ogive approaches at its extremes, but never quite reaches. The upper asymptote is located on the vertical axis at 1.00. In Figures 4.4 and 5.5 you can see that the upper, right part of the logistic ogives approach the value of 1.00 on the vertical axis. In the figures it may appear as though they touch the horizontal line at 1.00, but, strictly speaking, they never quite do.

5.9 The lower asymptotes for the ogives in Figures 4.4 and 5.5 is the horizontal axis with a height of zero. Just as the upper part of the ogive never quite reaches 1.00, the lower part of the ogive never quite reaches the lower asymptote.

5.10 All logistic ogives in IRT have an upper asymptote at 1.00, but not all have a lower asymptote at .00. In fact, few do.

5.11 Figure 5.11 shows 3 logistic ogives, labeled H, J, and K, which are identical except for different lower asymptotes. The lower asymptotes are at .15, .25, and .30 on the vertical axis. The b-value for each ogive = 0.0. Note that the upper asymptote for all 3 ogives is at 1.00.

5.12 Note also that the inflection points (all located at 0.0 on the θ scale) for the ogives in Figure 5.11 are at different heights. In fact, they are half-way between their asymptotes. That is always the case. The inflection point of the logistic ogive is always half-way between its upper and lower asymptotes.

5.13 The lower asymptote is called the c-parameter or the c-value. It is another of the 3 parameters of IRT.

5.14 The effect of the c-value is to squeeze the ogive into a smaller vertical range. The reduced range is equal to 1 - c. The effect of the reduced vertical range is to reduce the slope of the ogive at every point on the θ scale, other things being equal. We include the c-parameter in the logistic function by multiplying by 1 - c, and adding c.

$$\Psi(\theta) = c + (1-c)\left[1 + e^{-1.7(\theta - b)}\right]^{-1}$$

which is the same as

$$\Psi(\theta) = c + (1-c)\left[1 + \exp(-1.7(\theta - b))\right]^{-1}$$

and

$$\Psi(\theta) = c + \frac{(1-c)}{\left[1 + e^{-1.7(\theta - b)}\right]}$$

The c-values of ogives H, J, and K in Figure 5.11 are .30, .25, and .15, respectively.

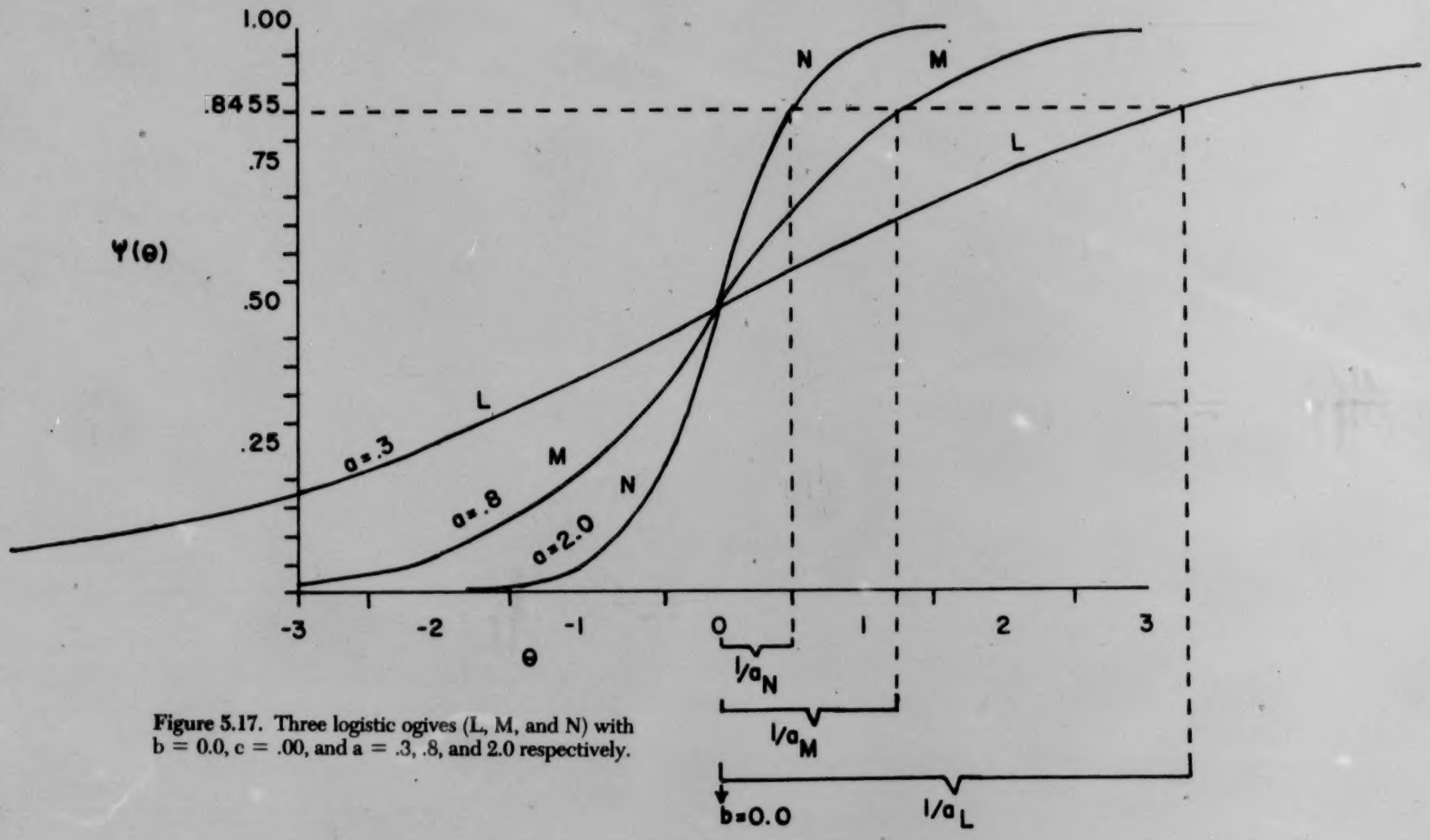


Figure 5.17. Three logistic ogives (L, M, and N) with b = 0.0, c = .00, and a = .3, .8, and 2.0 respectively.

5.15 The third (and last) parameter of IRT is (you guessed it) the a-parameter, or a-value.

5.16 The a-parameter is related to the slope of either ogive at the inflection point or in other words at the b-value. For the normal ogive model (with c = 0.0)

$$a = \sqrt{2\pi}\, m \approx 2.5m$$

where m is the slope of the ogive at the b-value.

5.17 Figure 5.17 shows 3 logistic ogives (L,M,&N), which are identical except for their a-values = .3, .8 and 2.0, respectively, with b = 0.0 and c = .00. As you can see, the larger the a-value, the steeper the ogive. Specifically,

$$a = \left[\Psi^{-1}(\theta) - b\right]^{-1}$$

where $\Psi^{-1}(\theta)$ = the point on θ, where the height of the ogive = c + .8455(1-c). The -1 that looks like an exponent of $\Psi(\theta)$ is not an exponent at all, but indicates the inverse of the function. Typically, a function is used by starting at some point on the abscissa, going vertically to the function, and then horizontally to the ordinate. The inverse procedure would be to start at a point on the ordinate (in this case at c + .8455(1-c)), go horizontally to the function, and then drop down to the abscissa (θ). That point on θ is $\Psi^{-1}(\theta)$. The -1 outside the brackets is an exponent, which means to take the reciprocal. The number .8455 is the proportion of area under the logistic f.f. and to the left of z-score = 1 (see Figure 4.1). The z-score = 1 is an arbitrary mathematically convenient point.

5.18 The a-parameter enters the logistic function as part of the exponent of e.

$$\Psi(\Theta) = c + \frac{1-c}{1+e^{-1.7a(\Theta-b)}}$$

This formula is the 3-parameter logistic ogive. It will look rather ominous to the novice. However, it is not difficult with a pocket calculator with an e^x key and a 1/x key. It is highly instructive to go through the calculation of several points of a typical logistic ogive and to plot them. An opportunity to do so is provided below for an ogive with a =.9, b = -.4, and c = .2. The reader can verify the results in Figure 5.18, which shows this logistic ogive with its characteristic parts labeled.

Pocket Calculator Instructions

a = .9

b = -.4 $\Psi(\theta) = c + \dfrac{(1-c)}{1+e^{-1.7a(\theta-b)}}$

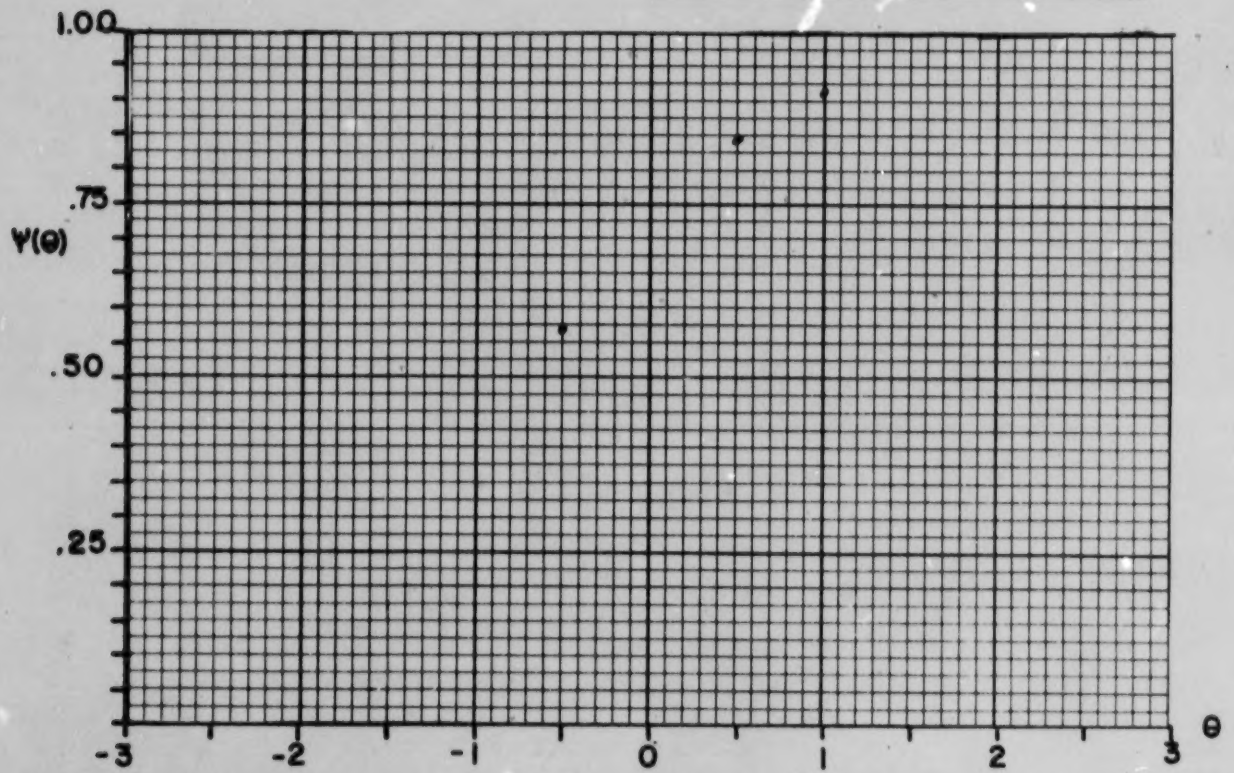
c = .2

Enter	Key	Comment
θ		(pick one)
	-	minus
-.4		b
	x	times
.9		a
	x	times
-1.7		constant
	=	-1.7a(θ-b)
	e^x	
	+	plus
1		constant
	=	
	1/X	reciprocal
	x	times
.8		1-c
	+	plus
.2		c
	=	$\Psi(\theta)$

Record your $\Psi(\theta)$ here

θ	$\Psi(\theta)$
3	____
2.5	____
2	____
1.5	____
1	.916
.5	.839
0	____
-.5	.569
-1	____
-1.5	____
-2	____
-2.5	____
-3	____

Now plot $\Psi(\theta)$ vs. θ below.



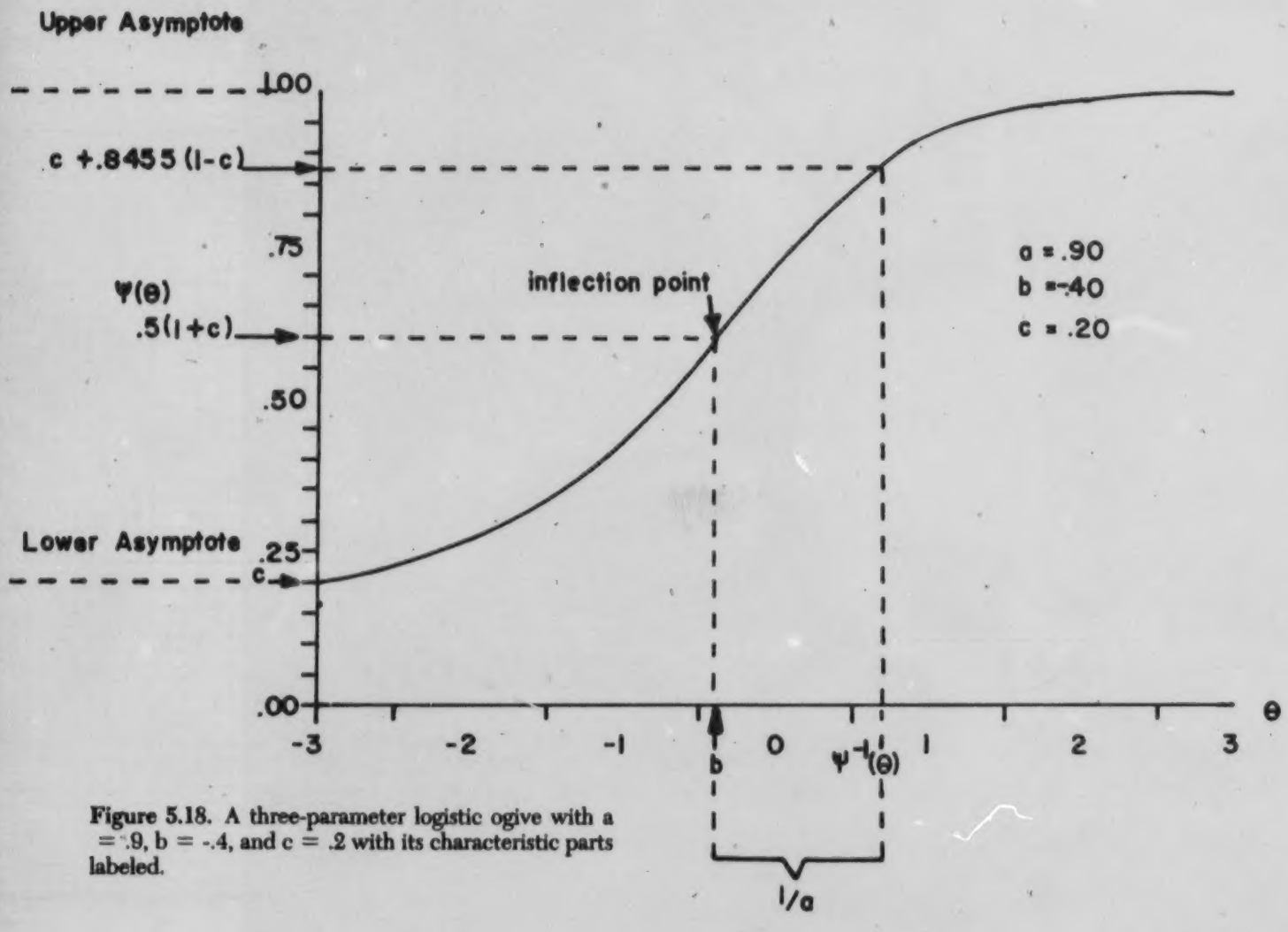


Figure 5.18. A three-parameter logistic ogive with a = .9, b = -.4, and c = .2 with its characteristic parts labeled.

CHAPTER 6

The Item Response Function (IRF)

6.1 Let's consider 2 examinees (Al and Bob) with different ability levels, i.e. different θs. Let's say Al has a higher θ than Bob. That means they are located at different places on the θ scale. See Figure 6.1.

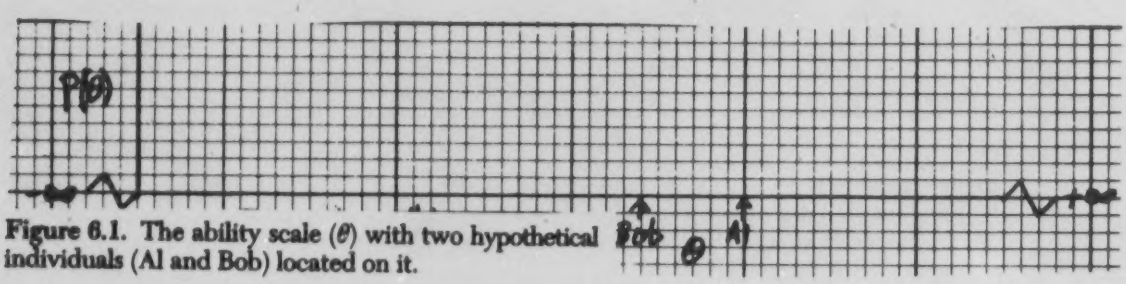


Figure 6.1. The ability scale (θ) with two hypothetical individuals (Al and Bob) located on it.

6.2 What are the chances that Al will get item #1 correct? What are the chances that Bob will get item #1 correct? So far we don't know the answer to either of those questions. But we do know one thing. Al has a better chance of getting item #1 correct than Bob, because Al is smarter than Bob (in ability θ). So let's represent the probability of each getting the item correct by a point above each (points A & B) in Figure 6.2.

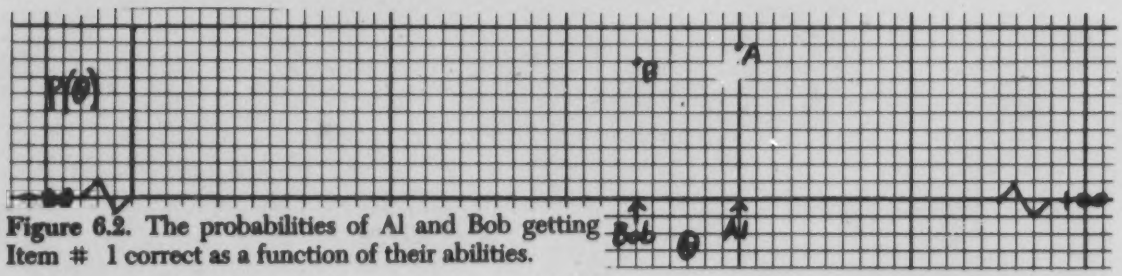


Figure 6.2. The probabilities of Al and Bob getting Item # 1 correct as a function of their abilities.

6.3 In doing so we have defined an ordinate as the probability of getting the item correct as a function of θ (ability). This may be written P_i (R|θ), and read, "the Probability of getting item i correct given (|) θ." But for brevity it is usually written $P_i(\theta)$. The subscript (i) is often omitted.

6.4 Now let's take Carl, who is dumber (less ability θ) than Bob. Carl has an even smaller chance of getting the item correct. See Figure 6.4a.

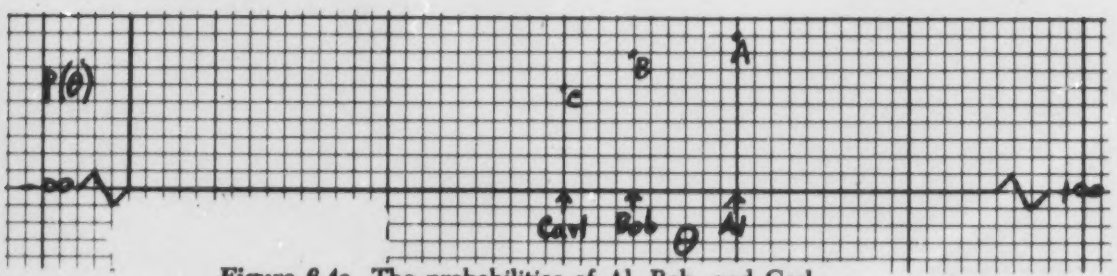


Figure 6.4a. The probabilities of Al, Bob, and Carl getting Item # 1 correct.

And let's also add Dave, and Ed and Fred who have less θ still. See Figure 6.4b.

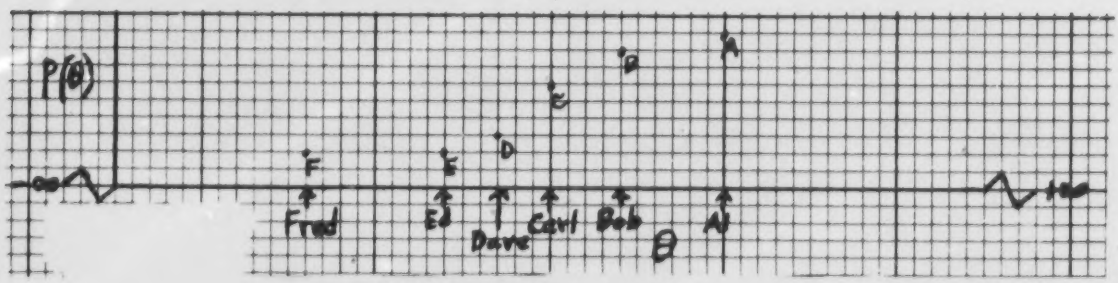


Figure 6.4b. The probabilities of Al, Bob, Carl, Dave, Ed, and Fred getting Item # 1 correct.

And we can add Olga, who is very bright. See Figure 6.4c.

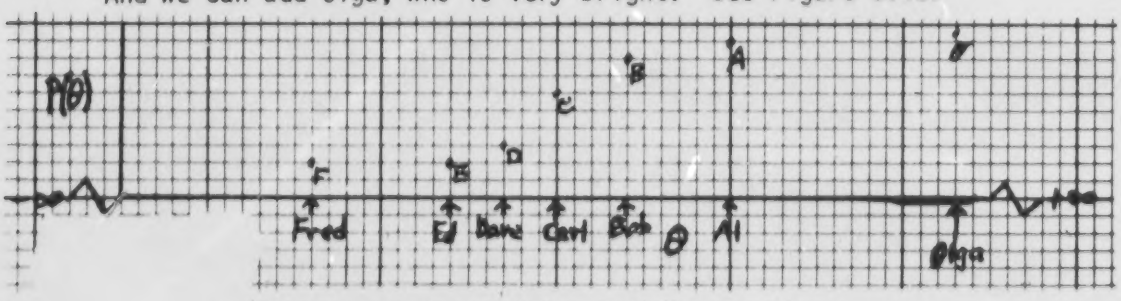


Figure 6.4c. The probabilities of Al, Bob, Carl, Dave, Ed, Fred, and Olga getting Item # 1 correct.

6.5 Since the probability of getting the item correct is only a function of the amount of ability,* we can say that any who has the same θ as Al will have the same probability as Al of getting the item correct (A). And, everyone who has the same θ as Ed will have the same probability as Ed of getting the item correct (E), and so on. Therefore, we can connect the points in Figure 6.4c, which will tell us the P(θ) for each θ. This curve is called the Item Response Function (IRF) and was until recently called the Item Characteristic Curve (ICC). See Figure 6.5

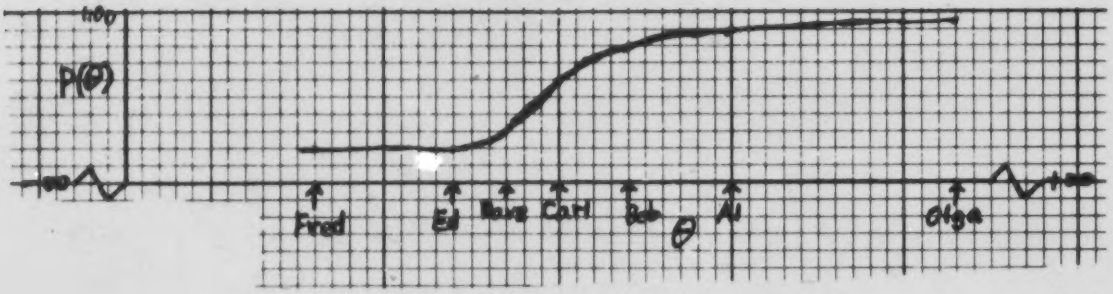


Figure 6.5. The Item Response Function of Item # 1.

6.6 We know several things about this IRF.

(1) It cannot rise higher than 1.0, because a probability = 1.0 is a sure thing, and nothing can be more probable than a sure thing.

(2) It will never reach a height of 1.0, because in testing there is no such thing as a sure thing. Therefore, the curve has an upper asymptote of 1.00.

(3) Between Ed and Bob the curve has to rise rapidly, because it must rise from point E to point B in the short distance between Ed's θ and Bob's θ.

*assuming unidimensionality, which will be discussed in Section 14.4.

(4) The curve must always rise (i.e. can never be horizontal or go down) as we move from left to right, because as ability increases, so does the probability of getting the item correct. Therefore, the curve is strictly monotonic.

(5) It cannot go below 0.00, because a probability = 0.00 is an absolute impossibility, and nothing can be less probable than an absolute impossibility. Therefore, the curve has a lower asymptote.

(6) Since the item is a multiple-choice question, there is usually a fair probability of getting the item correct strictly by chance alone, no matter how low the θ. Traditionally, we have taken this probability to be 1/A, where A = the number of alternatives in the multiple-choice question. A 4-choice item has been thought to have a chance probability of 1/4 = .25, and a 5-choice item, a chance probability of 1/5 = .20. Whatever the chance probability of getting a multiple-choice item correct is, it is not expected to be zero. It is expected to be somewhat greater than zero. Therefore, the curve in Figure 6.5 is expected to have a lower asymptote above zero. (In Section 7.3 we shall see that the lower asymptote is seldom 1/A)

6.7 You have probably noticed that all of the things we observed about the IRF are also true about the 3-parameter normal ogive and logistic ogive.

Therefore, we conclude that the normal (or logistic) ogive may be used to describe the IRF very well. And we may use the logistic ogive function to describe the IRF mathematically.

6.8 If somehow we knew and we were to plot the probabilities of getting item #2 correct for Al, Bob, Carl, Dave, Ed, Fred, and Olga, we might get an IRF like Figure 6.8.

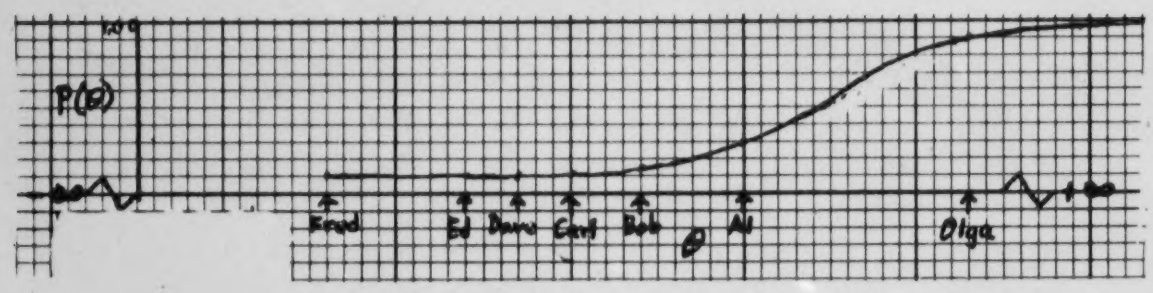


Figure 6.8. The Item Response Function of Item # 2.

6.9 Figure 6.9 shows both item #1 and item #2.

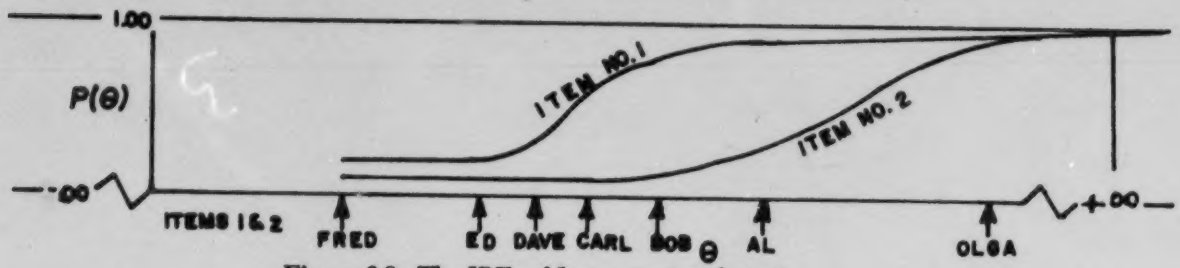


Figure 6.9. The IRFs of Items # 1 and # 2.

For Olga, Ed and Fred (and anyone else with their θs) the probability ($P_2(\theta)$) of getting item 2 correct is about the same as their $P_1(\theta)$ for item #1.

But item #2 is harder for Al, Bob, Carl, and Dave than item #1, because for all of them the probability of getting item #2 correct ($P_2(\theta)$) is lower than the probability of getting item #1 correct. And it would be harder for anyone who has the same ability as Al, Bob, Carl, or Dave.

6.10 We also notice that the probabilities of getting item #2 correct for Bob, Carl, Dave, Ed and Fred are all about the same. Item #2, then, does not do a good job in distinguishing among people with abilities like Bob's or below. This observation is consistent with what we intuitively understand about items. A hard item does not discriminate among low ability people, because they all get it wrong (unless they make a lucky guess). An easy item does not distinguish among high ability people, because they all get it correct. A test composed of items with IRFs like item #2's IRF would not be a good test for measuring the relative ability of people like Bob, Carl, Dave, Ed and Fred.

Note: In practice, any particular examinee may either know the answer to a particular item (in which case his probability of getting it correct is 1.00), or not know it (in which case his probability of getting it correct is chance). Strictly speaking, we can not talk about the probability of a particular person getting correct a particular item. However, for pedagogical reasons we will violate this restriction in this section.(See Section 8.2 for clarification.)

6.11 However, Olga's $P_2(\theta)$ for item #2 is much higher than Al's $P_2(\theta)$. Therefore, item #2 will distinguish between people like Al and Olga. If a distinction in that range of ability is our purpose, then a test made of items like #2 would be a pretty good test.

6.12 Item #3 might have an IRF like that in Figure 6.12. This item rises over a longer range than does either item #1 or item #2, but its slope is less at every point during its rise. This low slope means that item #3 is discriminating over a wide range of θ, but is not doing so well at any particular θ.

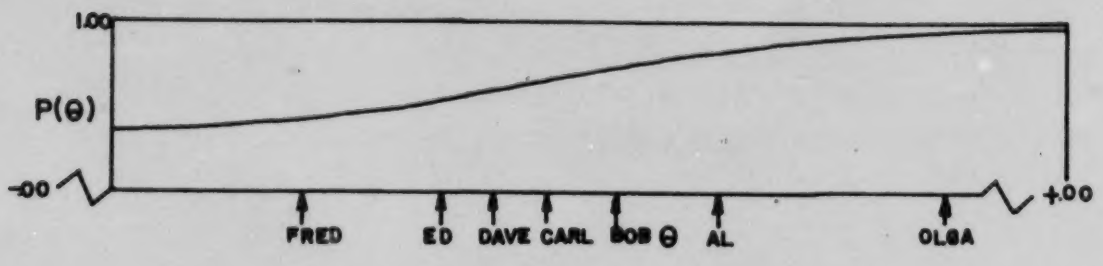


Figure 6.12. The IRF of Item # 3.

6.13 Figure 6.13 shows the IRFs for both item #1 and item #3.

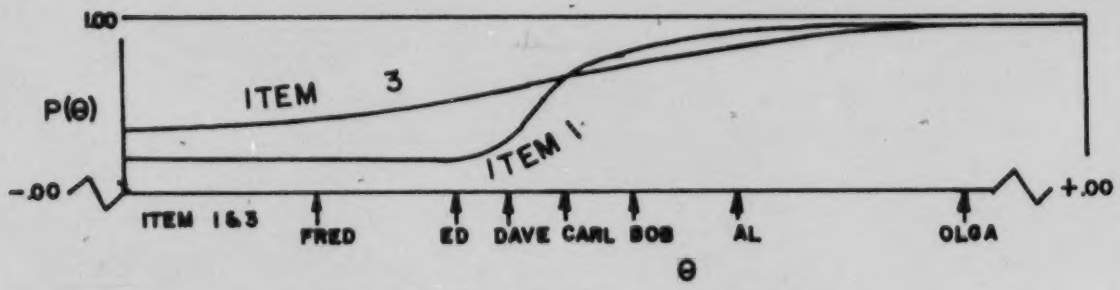


Figure 6.13. The IRF of Items # 1 and # 3.

46

It is interesting to note that item #3 is harder than item #1 for Al and Bob, but easier for Dave, Ed, and Fred. This possibility of reversed relative item difficulty for persons of different ability is one of the surprising results of IRT.

6.14 We have seen that the greater the slope of the IRF, the greater the discrimination, but the smaller the range of discrimination. We have already noted in Chapter 5 that the a-parameter of the logistic ogive describes its slope. Therefore, the a-value is called the discrimination index of the IRF. The greater the a-value of the IRF, the better the item discriminates.

6.15 Also apparent is the fact that the shift of the IRF as a whole to the left makes the item easier in general, and to the right makes the item harder in general. The left-right shift of the logistic ogive is described by the b-parameter. Thus, the b-value is the difficulty index of the IRF. The more difficult the item is, the larger (in the positive direction) the b-value of the IRF.

6.16 The IRFs of items 1, 2, and 3 have different lower asymptotes. Since the IRF never goes below the lower asymptote, this difference in IRFs means that the items are of different difficulty even for examinees of very low ability. But examinees of very low ability will know almost nothing about the item, and therefore have to guess. The difference in lower asymptotes of IRF's means that very low ability examinees have a better chance of guessing the correct choice of some items than of others. This result of IRT will be discussed further in Section 7.3. The lower asymptote of the logistic ogive is the c-parameter. The c-value of an IRF is called the "guessing index" or more properly the "pseudo-guessing index" of the item. Both terms are used.

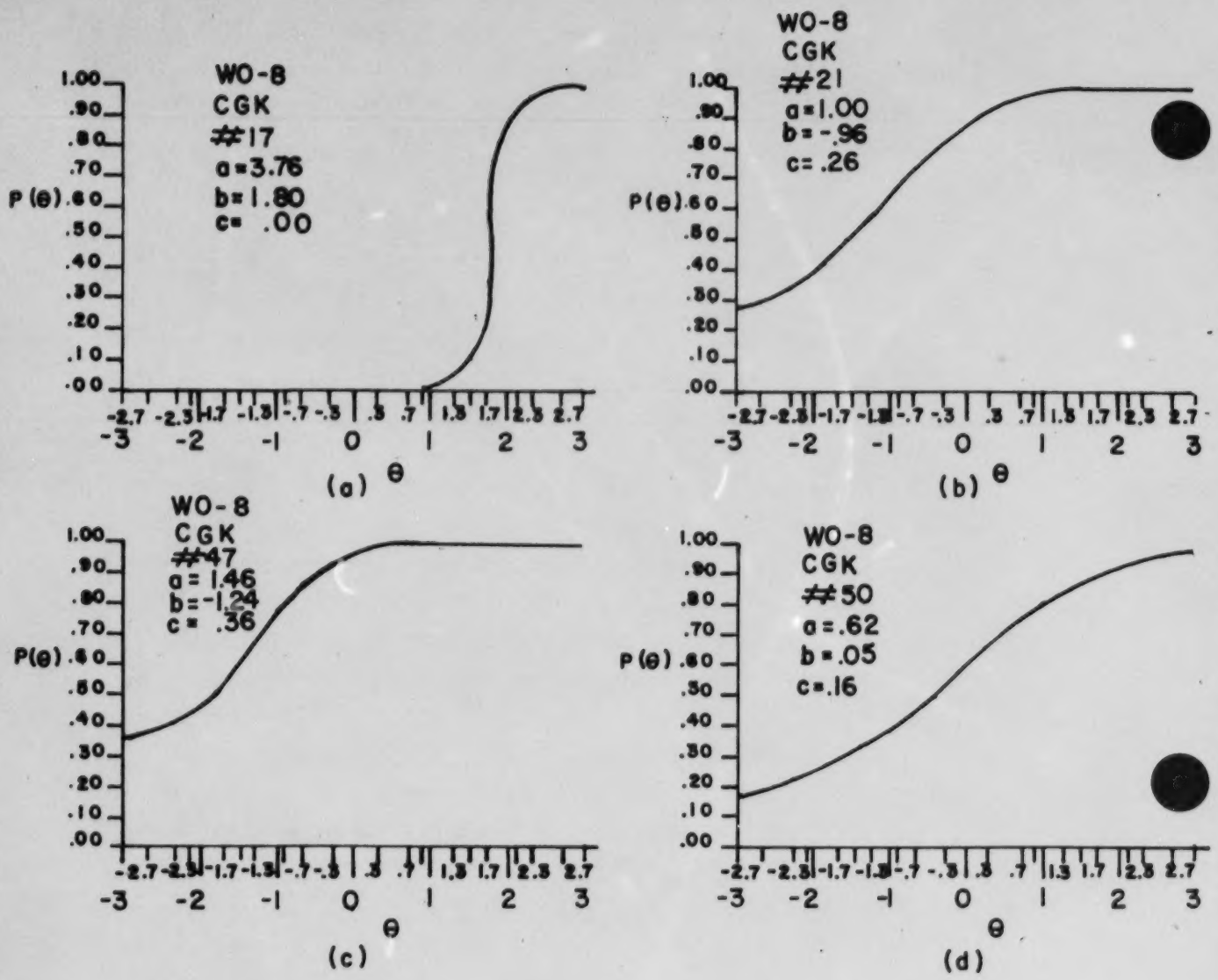


Figure 6.17. The IRFs of four actual items from the Coast Guard Knowledge section of the U. S. Coast Guard Warrant Officer Test, series 8.

48

6.17 Figure 6.17 shows the IRF's for 4 actual items from the Coast Guard Knowledge section of the U.S. Coast Guard Warrant Officer test. Item #17 is a very difficult, but highly discriminating item. It has a c-value of .00, which means that nearly all examinees below $\theta = 1$, answered the item incorrectly. Item #17 is a very unusual item in two respects, its extremely high a-value, and .00 c-value. It is, however, an ideal item for many purposes.

Item #21 is an easy item with somewhat low discrimination. Item #47 is slightly easier than #21, but has good discrimination. Item #50 is an item with medium difficulty, and poor discrimination.

6.18 The IRF should not be confused with the item-test curve. The item-test curve has raw score as the horizontal axis instead of θ. The item-test curve, therefore, suffers from the same problems of distorted scale as the raw score. The item-test curve has no particular shape, and is not independent of the other items in the test. In fact, the average of the item-test curves of all items in a test is always a straight line of slope = 1(i.e. 45°). Thus, for many purposes the item-test curve is useless as an analytic tool.

CHAPTER 7

The a, b, & c parameters

7.1 The a-value is the discrimination index of the item. If θ is normally distributed, in the normal ogive model the a-value is related to the d-value in the following very complex way (from Schmidt, 1977).

$$a \approx \frac{d\sqrt{pq}}{\sqrt{(KR\text{-}20)(1-c)^2y^2 - d^2pq}}$$

where d = d-value, the point biserial item-test correlation

p = p-value, the proportion of examinees correctly answering the item

q = 1-p

KR-20 = Kuder-Richardson formula 20 reliability

y = the height of the N(0,1) curve at the z score that cuts off p' proportion of the area under the N (0,1) frequency function.

c = c-value

$$p' = \frac{p-c}{1-c}$$

The a-value is related to the slope of the IRF, and can range from 0.0 to + ∞ just as the slope can. Negative slopes are possible, but not of interest to us. Experience has shown that a-values of typical items vary from about .5 to 2.5 with most from 1.0 to 2.0. The highest I have observed is 3.76. An item with a low a-value discriminates poorly over a wide range of θ. With a high a-value the item discriminates well, but over a small range of θ. Items with a-values below .80 are not very good items for most purposes.

7.2 The b-value is the difficulty index. If θ is normally distributed, it is related to the p-value in the normal ogive model (from Schmidt, 1977) in the following way:

$$b \approx \frac{yz(1-c)\sqrt{KR\text{-}20}}{d\sqrt{pq}}$$

where z = the z-score that cuts off p' proportion in the upper portion of the area under the N(0,1) frequency function, and the other symbols are as defined in Section 7.1 above. Typical b-values range from -2.5 to +2.5. A b-value of -2.5 indicates the item is very easy. An item with a +2.5 b-value is very difficult, and items with 0.0 b-values are of medium difficulty.

7.3 The c-value is the guessing parameter or pseudo-guessing parameter. It indicates the probability of examinees with very low ability of getting the item correct. Most c-values range from .00 to .40. Items with c-values of .30 or greater are not very good items. It is desirable to have the c-value at .20 or less. The lower the c-value is, the better. A zero c-value is ideal. Typically, the c-value is about 1/A - .05, where A = the # of alternatives. Thus, 4-choice items often have c $\approx$ .20 (i.e. .25-.05), and 5-choice items often have c $\approx$ .15 (i.e. .20-.05).

Items do not have a c-value of 1/A because examinees do not, in fact, guess randomly when they do not know the answer (as has often been assumed in classical test theory analyses).

7.4 Two explanations have been offered for the fact of non-random guessing (c $\neq$ 1/A).

Lord has suggested that item writers are very clever in writing distractors that are very attractive to low ability examinees. Thus, when low θ examinees do not know the answer they are attracted more to distractors than to the correct answer, and so get the item wrong more often than if they guessed randomly.

The other explanation is my own, based upon personal knowledge of item writing and test taking behavior:

(1) When an item writer sits down to write items, he, for the moment, is not concerned with the distribution of the correct answers (the keyed choices) among the four (for four-choice items) possible positions (i.e. choice A, choice B, choice C, and choice D).

(2) He has a tendency to try to hide the correct choice. In a four-choice item there are only 2 places to hide it - choice B, or choice C. Therefore, he writes many more items, keyed B or C than A or D, and in fact there seems to be a much stronger tendency toward C. (I have verified this tendency with many item writers). This also seems to be true for 5-choice items.

(3) When he finishes writing the items, he tabulates the numbers of items keyed for each position, and usually finds that he has many more C's than A's, B's, or D's (or E's in 5-choice items).

(4) Most testing organizations have a requirement that there should be about equal numbers of items with the keyed choice in each of the 4 or 5 possible positions.

(5) The item writer then begins to revise the order of the choices in items to decrease the number of items keyed C, and increase the number of items keyed A and D and maybe B. He continues to revise the order of the choices of items until he has satisfied the requirement of about equal numbers of keyed choices in each position.

(6) Naturally, to save himself work and time (the Law of Least Effort) he wants to revise as few items as possible. Therefore, he stops revising items when he gets within the requirement of about equal numbers. Because he started with more items keyed C, he also ends up with more items keyed C (but not as many), because he only needs about equal numbers.

If the above scenario is as universal as I believe, it means that, in the set of all multiple-choice items in the world, more are keyed C than any other choice. It is true of almost all of the tests I have checked.

There is a widespread rule of thumb among examinees: "If you don't know at all, guess C." I have heard this rule of thumb from coast to coast, from high school and college students, and from civilian employees and military personnel taking promotional tests. I do not know the source of this rule of thumb, but it is possible that the rule of thumb gradually grew from examinees' observations of the frequency of keyed choice positions, as I have suggested above.

Whatever the origin of the rule of thumb, it represents rational behavior, given a higher frequency of choices, keyed C, among the population of all multiple-choice items. By choosing choice C (when you don't know at all), you will get more items correct by chance in the long run than by guessing at random.

This analysis suggests that the c-values of items keyed C will be higher than for items keyed A, B, and D. I was able to test this hypothesis with 127 items from 6 forms of the verbal parts of the SCAT-II series of tests, published by the Educational Testing Services, Princeton, NJ. The c-values were provided by Fred Lord. A two-by-two frequency table of A, B, D vs C by above-average c-value vs below-average c-value yielded a Chi square significant beyond the .001 level. This result strongly supports the hypothesis that low ability examinees get items keyed C correct more often than they get items keyed A, B, or D correct.

The results suggest 2 alternative courses of action for testing organizations.

(1) Require that there be exactly the same number of keys in each position. This action would thwart the test-wiseness of those who use the rule of thumb. However, it represents an undesirable rigidity.

(2) A better course of action would be to key C for less than 1/4 of the items (for 4-choice items). This action would cause a lower average c-value for the test. The lower average c-value would increase the total information in the test, which as we will see in Sec. 9.4 is highly desirable.

7.5 The Rasch model assumes that all items in a test have the same a-value, and that c = .00 for all items. Both assumptions are nearly always unrealistic.

CHAPTER 8
The Test Characteristic Curve

8.1 The scale of θ is continuous, but since most of the calculations are done on digital computers, θ is usually broken into small, discrete intervals of .05 θ units, and values of $P(\theta)$ are calculated for each .05 interval from $\theta = -5.0$ to $\theta = +5.0$. The very broad range from -5.0 to 5.0, and the small .05 intervals are used in the interest of accuracy. Larger or smaller intervals and a broader or narrower range may be used depending on the purpose and degree of accuracy desired.

8.2 Table 8.2 below gives the $P(\theta)$ for 17 values of θ for each of the 4 items, shown in Figure 6.17.

P(θ)

θ	#17	#21	#47	#50	$\Sigma P(\theta)$
-2.7	.00	.30	.38	.20	.88
-2.3	.00	.33	.40	.23	.96
-2.0	.00	.37	.45	.25	1.07
-1.7	.00	.43	.52	.28	1.23
-1.3	.00	.53	.66	.33	1.52
-1.0	.00	.71	.87	.44	2.02
-.7	.00	.62	.77	.48	1.77
-.3	.00	.82	.94	.52	2.28
0	.00	.88	.97	.59	2.44
.3	.00	.92	.99	.65	2.56
.7	.00	.96	.99	.74	2.69
1.0	.01	.97	.99	.79	2.75
1.3	.04	.98	.99	.84	2.85
1.7	.35	.99	.99	.89	3.22
2.0	.78	.99	.99	.91	3.67
2.3	.96	.99	.99	.94	3.88
2.7	.99	.99	.99	.96	3.93

Table 8.2

An item is scored dichotomously, which means the examinee either gets the item correct (for which he gets an observed score of 1) or he gets the item wrong (for which he gets an observed score of 0). The dichotomous score is a result of the typical use of multiple-choice items. An examinee's dichotomous score (0 or 1) is not a very accurate measure of his knowledge.

$P(\theta)$ may be interpreted in two ways. A $P(\theta) = .78$ means both:

(1) 78% of the examinees with the given θ will get the item correct, and

(2) An examinee will get correct 78% of the items for which his $P(\theta) = .78$.

If an examinee answers 100 questions for all of which his $P(\theta) = .78$, he is expected to get 78 items correct and 22 items wrong for a % score of 78%. If there were some way to give him partial credit of .78 points for each of the 100 items instead of 0 or 1 point he would also get a % score of 78%. This notion of partial credit for an item depending on his $P(\theta)$, leads to the idea of a true score on the item.

It is often not true that the examinee is 100% or 0% certain of his answer. Yet on a multiple-choice item he either gets full (100%) credit for the item (1, if he gets it correct) or no (0%) credit (0, if he gets it wrong). The examinee's degree of certainty, if measurable could be taken as a more precise measure of his knowledge. $P(\theta)$ might be interpreted as this measure of his knowledge, and is called his true score on the item. The sum of his true item scores is his true test score. His true test score is the raw score he would get, if there were no measurement error in the test.

The far right column in Table 8.2 is the sum of the $P(\theta)$'s of the 4 items for each of the listed points on the θ scale. The $\sum P(\theta)$ is the true test score of an examinee with a given θ on a test composed of the 4 items.

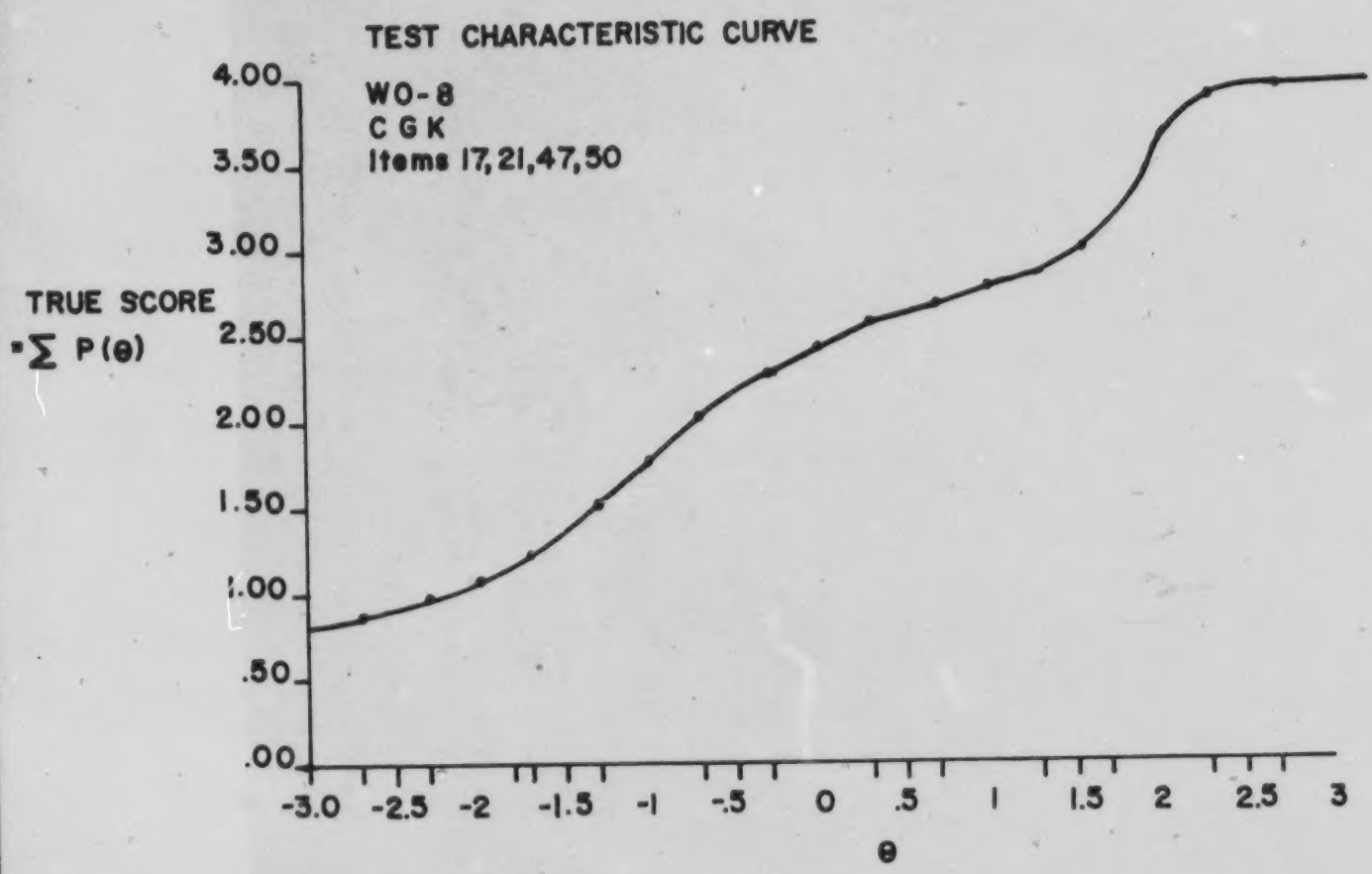


Figure 8.3. The Test Characteristic Curve of a test composed of four real items.

8.3 If we plot the true test scores against θ, we get a test characteristic curve (TCC). Figure 8.3 shows the TCC. The TCC gives the true score for each point on the θ scale. Notice that the TCC is neither a straight line nor an ogive. Each test will have its own TCC, which is the sum of the IRF's of the items in the test.

8.4 One of the interesting uses of the TCC is to determine the distribution of the true scores on the test. Figure 8.4 shows how this is done. If the examinees' θs are normally distributed, as shown on θ (upside down), the examinees' true scores will be as shown on the left. The true score distribution is found by projecting the intervals from the θ scale onto the TCC, and then representing the same area on the true score scale within the projected intervals. Figure 8.4 is an excellent demonstration of how the peculiarities of a test produce a distorted metric.

8.5 It is important to note that true scores (T) are not observed scores (X). Observed score is defined as true score plus error ($X = T + E$). However, Lord (1969) has found that the distribution of X will be similar to the distribution of T, but sometimes with the high points of the true score distribution flattened somewhat, and the low points higher. The flattening is due to error.

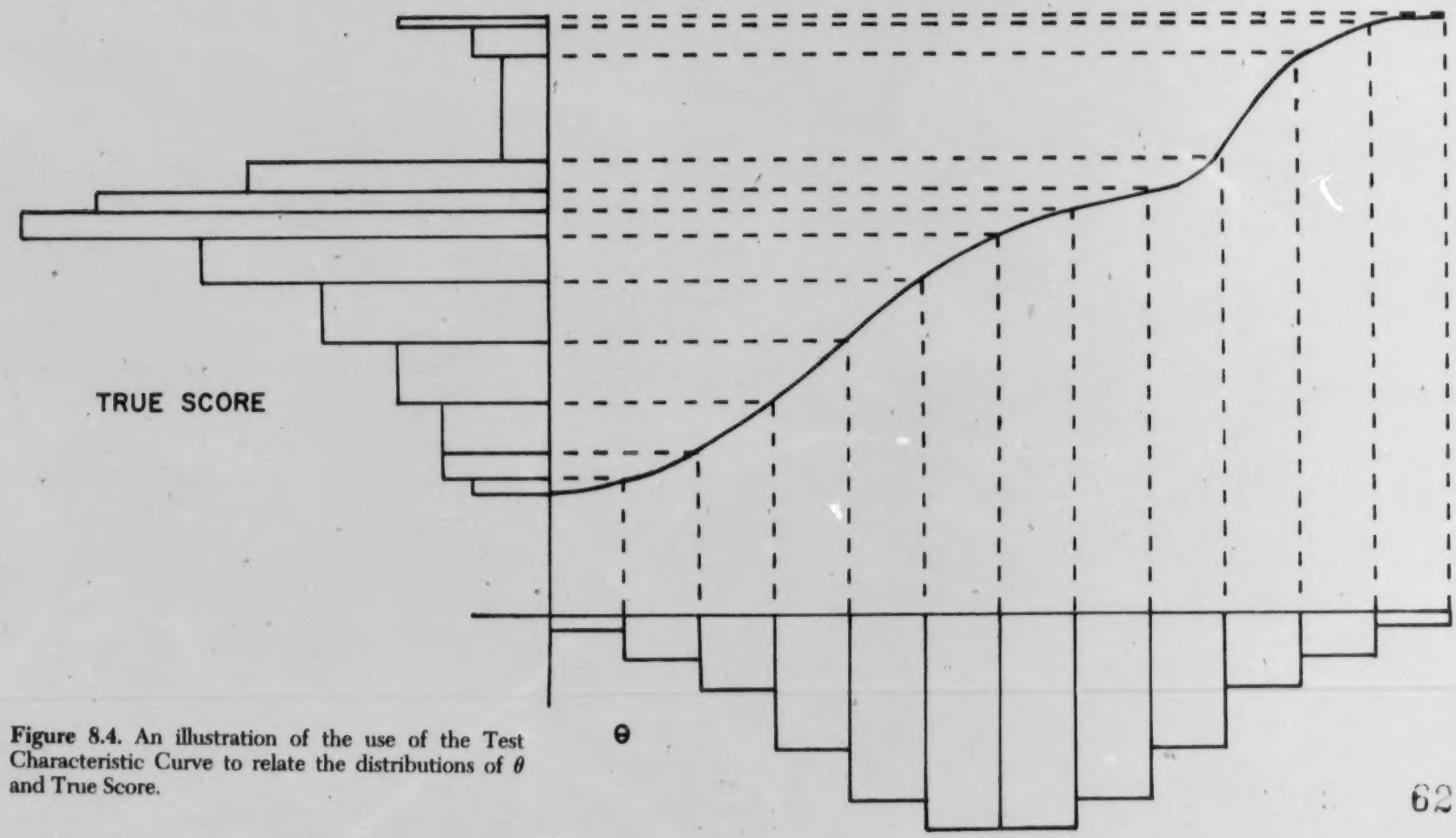


Figure 8.4. An illustration of the use of the Test Characteristic Curve to relate the distributions of θ and True Score.

CHAPTER 9
The Item Information Function (IIF)

9.1 We can see in Figure 6.17a that item #17 will not help us to distinguish among examinees whose θ's are less than 1.0 because they will all get the item wrong. Apparantly, there is something about item #17 that leads all examinees with $\theta < 1.0$ to choose the wrong alternative. This is an unusual situation, but actually occurs with this question. A test made exclusively of items like #17 would do nothing to distinguish among examinees with $\theta < 1.0$ because they would all get zero on the test. It would give us no distinguishing information about them.

Item #17 also gives us no distinguishing information about examinees with $\theta = 2.7$ or greater because they will all get it correct. On a test composed of items like #17, all examinees with $\theta > 2.7$ would get 100%.

Between $\theta=1.0$ and $\theta=2.7$, it is a different story. From $\theta=1.0$ to $\theta=1.5$, $P(\theta)$ goes from $P(\theta=1.0)=.00$ to $P(\theta=1.5)=.08$. The change of $P(\theta)$ means that the item does help to distinguish among examinees within the range of θ where the change of $P(\theta)$ occurs. In this case the difference between the $P(\theta)$'s (to be denoted dp) = .08 (.08-.00) is small. The change (dp) occurs over a range ($d\theta$) of 1/2 θ units (1.5-1.0). The ratio of dp to $d\theta$ ($dp/d\theta$) is equal to the average slope of the IRF over the range of $d\theta$. For the range from $\theta=1.0$ to $\theta=1.5$, $dp/d\theta = .08/.5 = .16$.

From $\theta = 1.5$ to $\theta = 2.0$ for item #17, $P(\theta)$ changes from .08 to .78, a very large change. $dp = .70$ (.78-.08) in this range, and $dp/d\theta = .70/.5 = 1.40$, which is very large. Item #17 is an excellent item for distinguishing among examinees in the range $\theta = 1.5$ to $\theta = 2.0$. A test composed of items like #17 would give scores from about 8% to 78% for examinees whose θ's go from 1.5 to 2.0. This test would give us a lot of distinguishing information about examinees in this range of θ, because it would spread them out over a wide range of test scores.

We can see that the greater the slope of the IRF, the more information the item gives us about examinees in the range being considered.

9.2 If we could make the range of θ over which we find the slope smaller and smaller, we would eventually get to the slope of the IRF at a point which would be the slope of the tangent line to the IRF at a particular point of θ.

The slope of the IRF would be a measure of the relative amount of information the item gives about examinees at that point. The greater the slope, the more information.

Fortunately, there is an easy way to find the slope of the logistic ogive. The slope of the IRF is given by:

$$p' = \frac{dP}{d\theta} = \frac{1.7a(1-c)e^{1.7a(\theta-b)}}{\left[1+e^{1.7a(\theta-b)}\right]^2}$$

where a, b, and c are the item parameters and θ is the point where dp/dθis the slope. The slope is also sometimes denoted as P'(θ), or P' for short. In calculus P'(θ) is known as the first derivative of P(θ). Since the slope (P') is a measure of information, it is possible to plot a curve that shows the amount of information an item gives at each point on the θ scale.

9.3 However, there is a catch. For mathematical and statistical reasons which we will not go into, P'(θ) is not a completely appropriate measure of information, but a related function is. The function is:

$$I(\theta,u) = \frac{p'^2}{P(\theta)Q(\theta)} = \frac{(1.7a)^2(1-c)}{\left[c+e^{1.7a(\theta-b)}\right]\left[1+e^{-1.7a(\theta-b)}\right]^2}$$

where P'^2 is P' squared, and Q(θ) = 1 - P(θ). Note that the exponent of the left e in the denominator is positive, and the exponent of the right e is negative.

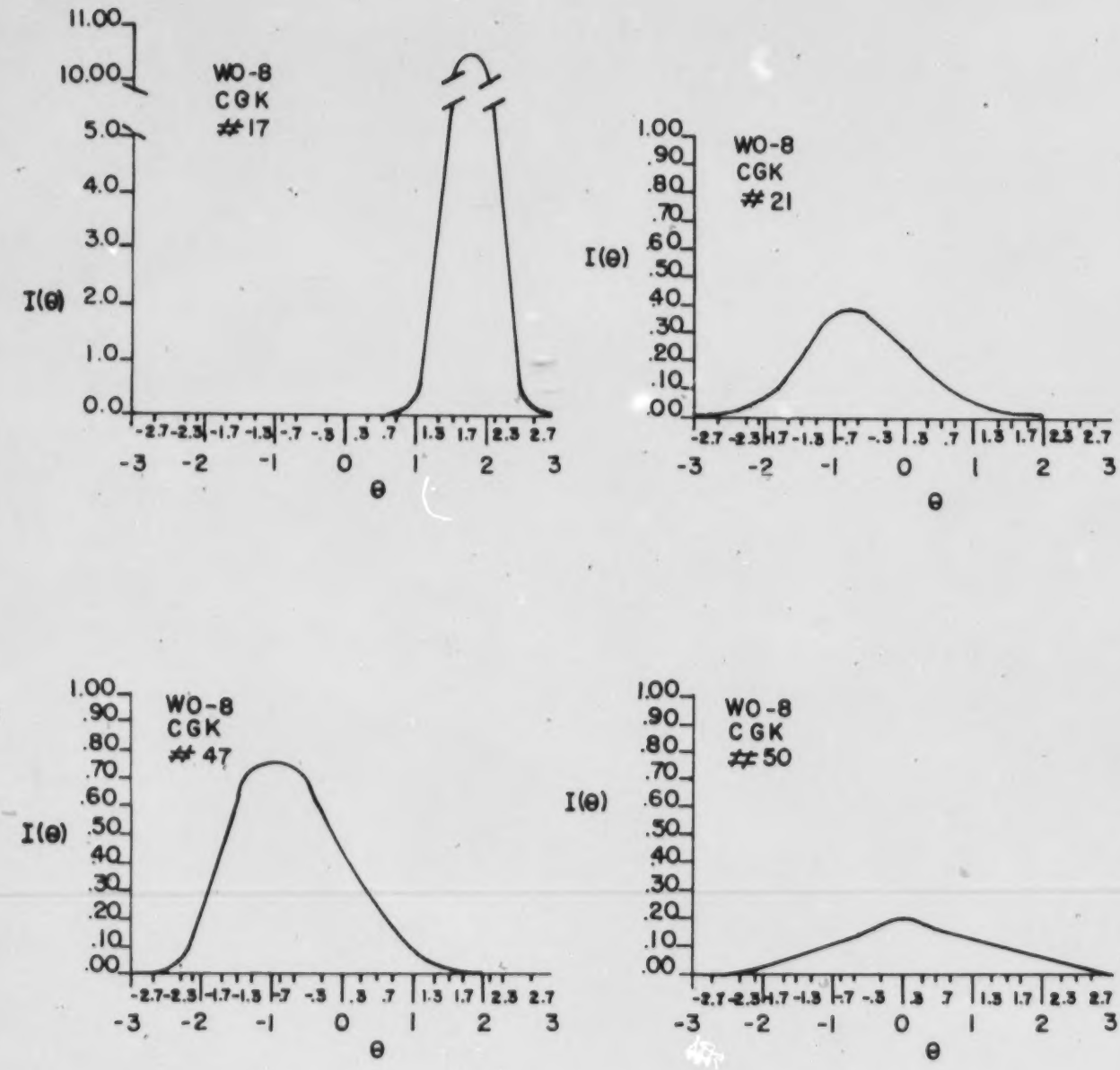


Figure 9.4a. The Item Information Functions of four real items.

That function is called the Item Information Function (IIF), and is written $I(\theta,u)$. The above formula for $I(\theta,u)$ may look even more ominous than the formula for $P(\theta)$, but in fact it is only slightly more complicated. It is still feasible to calculate points of $I(\theta,u)$ with a typical scientific hand calculator.

9.4 Figure 9.4a shows the $I(\theta,u)$ for the four items whose IRF's are shown in Figure 6.17. (Note that the vertical scale for item #17 is different from the others.) In comparing the IRFs with the IIFs, you will note three important relationships.

(1) The IIF is highest close to where the slope of the IRF is steepest.

(2) The total area under the IIF increases as the a-value increases.

(3) The total area under the IIF decreases as the c-value increases.

The fact that total information (i.e. total area under the IIF) increases as the a-value increases, demonstrates the importance of high a-values for items. However, there is another effect of high a-values. As the a-value increases, the width of the θ scale over which the information is distributed decreases. The effect is called the bandwidth paradox*. Thus, sometimes a compromise must be made between the total information provided by the item and the distribution of information over θ.

*This bandwidth paradox is different from the bandwidth paradox described by Cronbach (1960, p.602).

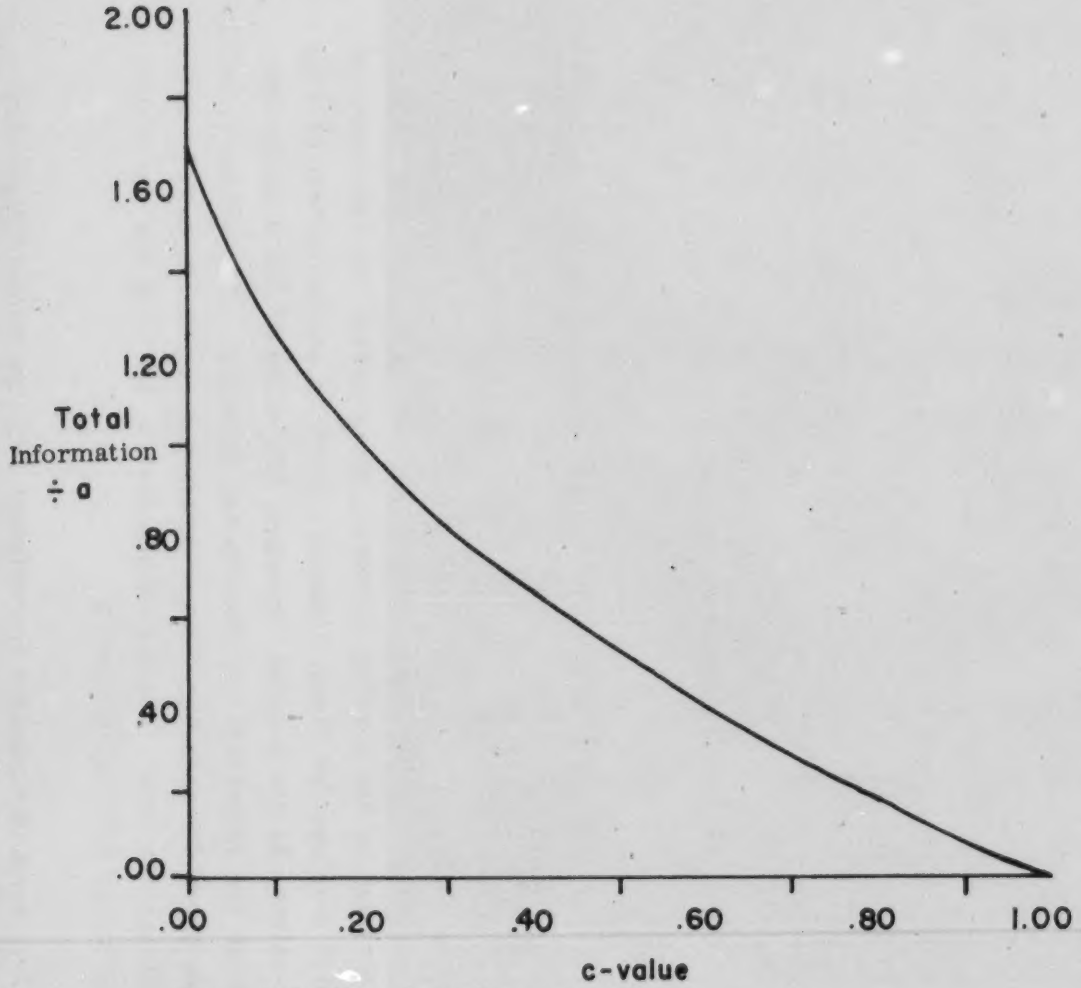


Figure 9.4b. The relationship of the c-value to the total information provided by an item (given a).

The total information (A_g) of item g is given by

$$Ag = \frac{1.7a\,(c \cdot \log c + (1-c))}{1-c} = \qquad 1.7a + \frac{1.7ac\log c}{1-c} = 1.7a\left(1 + \frac{c\log c}{1-c}\right)$$

where a and c are the item parameters and log c is the natural logarithm of c. From inspection of the formula for A_g, you can see that as the a-value increases, so does A_g. Also apparent is the fact that, as c approaches zero, A_g approaches 1.7a. Therefore, the maximum total information an item can provide is 1.7a. Not so obvious from the formula for A_g is the relation that, as c approaches 1.00, A_g approaches zero. This occurs because log c is negative except when c = 1, and because when c = 1, c log c/(1-c) = -1. This relation explains the effect of the c-value: the c-value destroys information. Figure 9.4b shows how total information decreases as c increases while holding the a-value constant.

Since the b-value is not included in formula for A_g, the b-value does not affect the total information.

9.5 The point on θ where the IIF is highest is not at the b-value, as one might expect (except when c=0). The point on θ where information is greatest is given by

$$\theta_{max I_{(\theta,u)}} = b + \frac{1}{1.7a}\left[\log\left(.5 + .5\sqrt{1+8c}\right)\right]$$

where "log" means the natural logarithm.

The point on θ where information is maximized is always to the right of the b-value, (except when c=0, it is at the b-value), but never farther to the right than .41/a.

9.6 The IIF is symmetrical when c=0 and skewed to the right when $c \neq 0$. The larger is c, the greater the right-skew. The right-skew occurs because the c-value destroys more information at low levels of θ than at high levels. This result makes sense because examinees at low θs will guess more than examinees at high θs. Guessing (i.e. the opportunity to get the item correct by guessing) destroys information. It is for this reason that five-choice items are preferred to four-choice items.

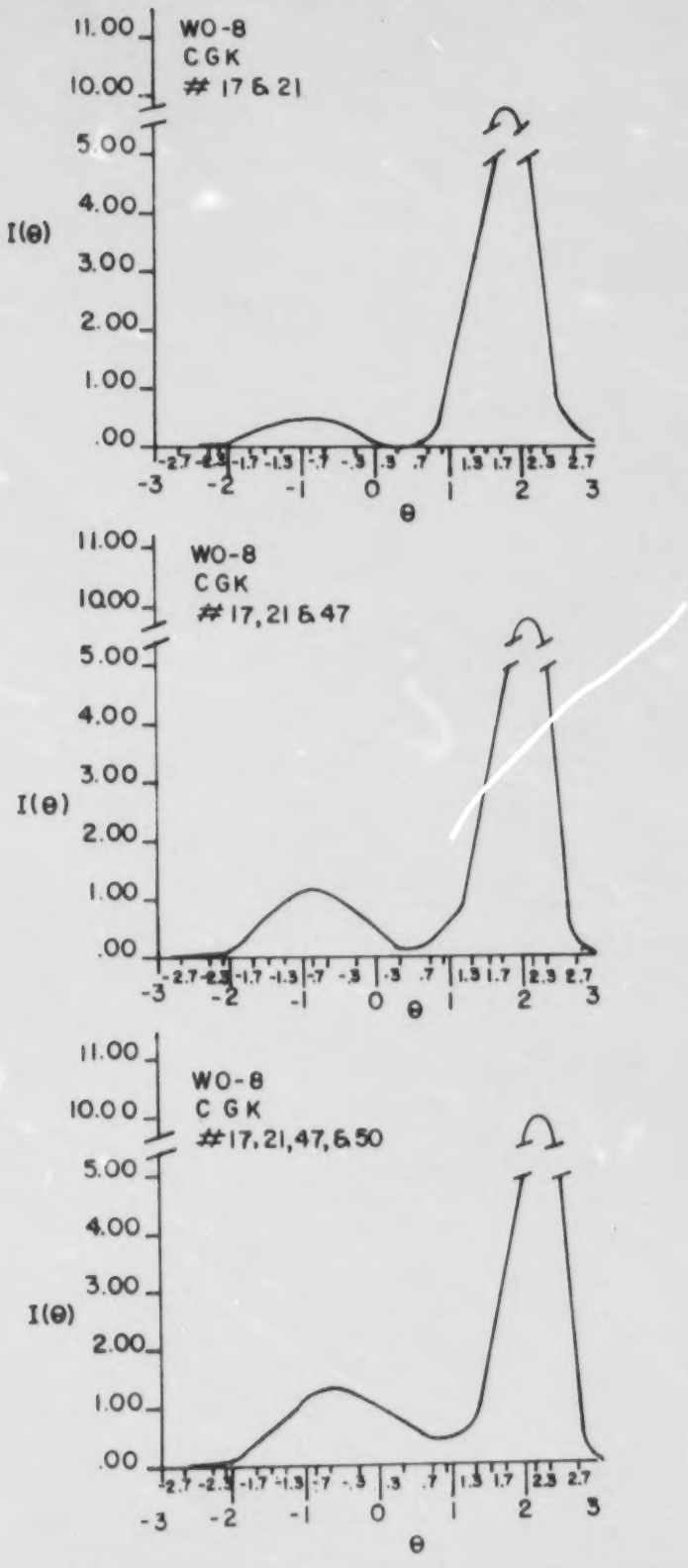


Figure 10.2 a, b, and c. The Test Information Curve of (10.2a) a test composed of items # 17 and # 21, (10.2b) a test composed of items #17, # 21, and # 47, and (10.2c) a test composed of items # 17, # 21, # 47, and # 50 from the USCG Warrant Officer Test.

CHAPTER 10

The Test Information Curve and Relative Efficiency Curve.

10.1 The Test Information Curve (TIC) is nothing more than the sum of the IIFs. IIFs are summed by "stacking them on top of each other." "Stacking" IIFs merely means that the heights (i.e. the amount of information) of the IIFs at a particular value of θ are added together to get the height of the TIC at that value of θ. Plotting the sum of item information at each value of θ gives the TIC. The height of the TIC at θ is written as $I(\theta)$.

$$I(\theta)=\Sigma I(\theta,u)$$

10.2 Figure 10.2a shows the sum of the IIFs for items #17 and 21 as shown in Figure 9.4a. Figure 10.2b shows the IIF of item #47 added to Figure 10.2a. Figure 10.2c shows the IIF of item #50 added to the other 3 items. A test composed of these four items would have the wierd TIC in Figure 10.2c.

10.3 The TIC shows the relative amounts of information provided by the test at each point on θ. Where you want information depends on what you will use the test for. If you want to select a few examinees from a large number, then you want a lot of information at high levels of θ, so that you can tell just which examinees are the best. For example, see Figure 10.3a. If you want to select all examinees except a few, then you want a lot of information at low θs so you can tell which examinees are the worst (e.g. see Figure 10.3b).

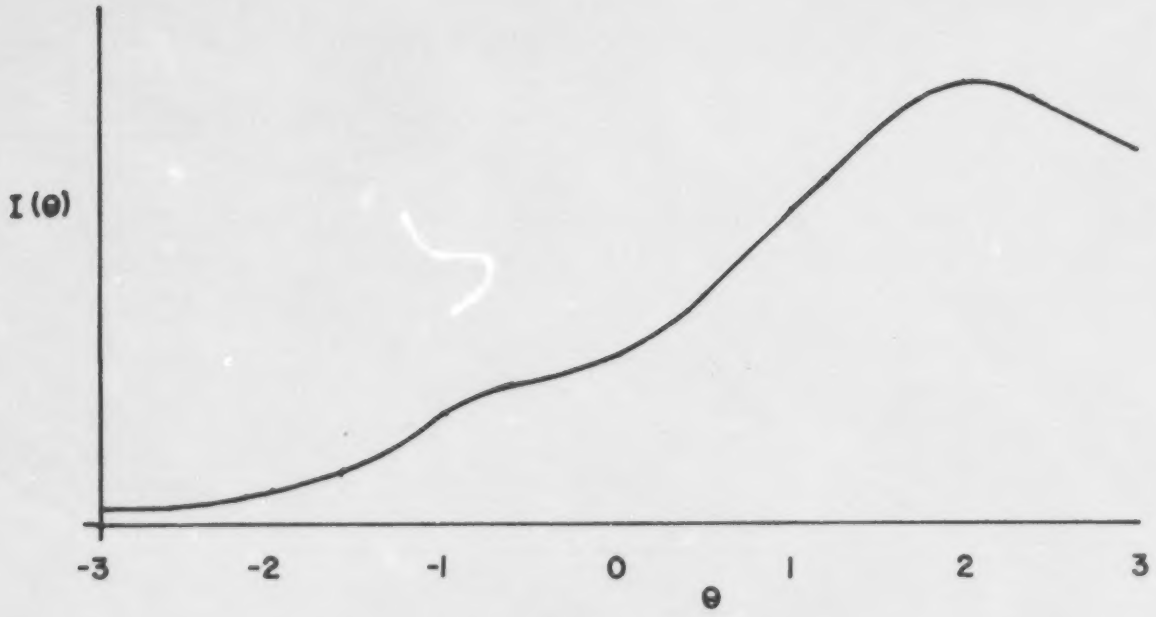


Figure 10.3a. Test Information Curve of a hypothetical test, which would be efficient for a high cut score ($\Theta = 2.0$).

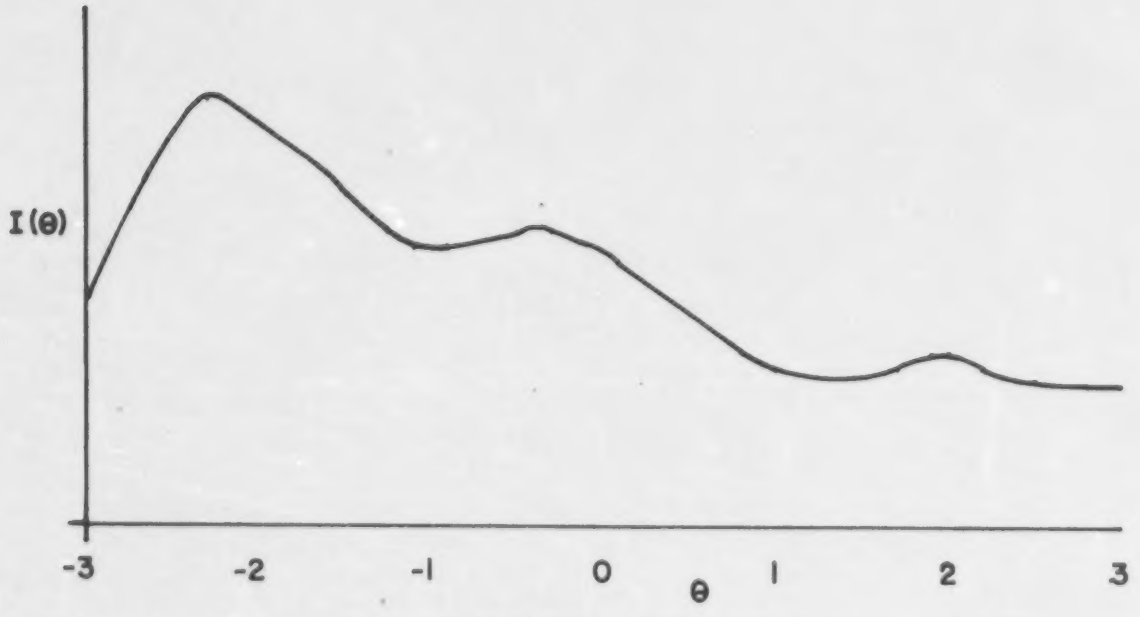
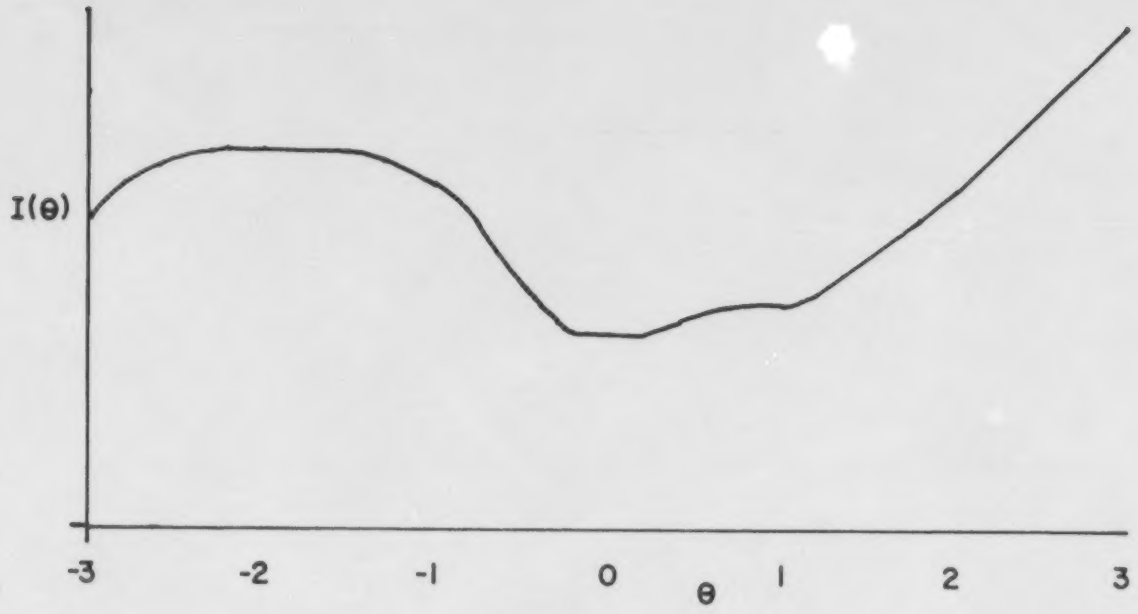


Figure 10.3b. Test Information Curve of a hypothetical test, which would be efficient for a low cut score ($\Theta = -2.3$).



Figures 10.3c The Test Information Curve of a hypothetical test, which would be efficient at both high and low cut-scores.

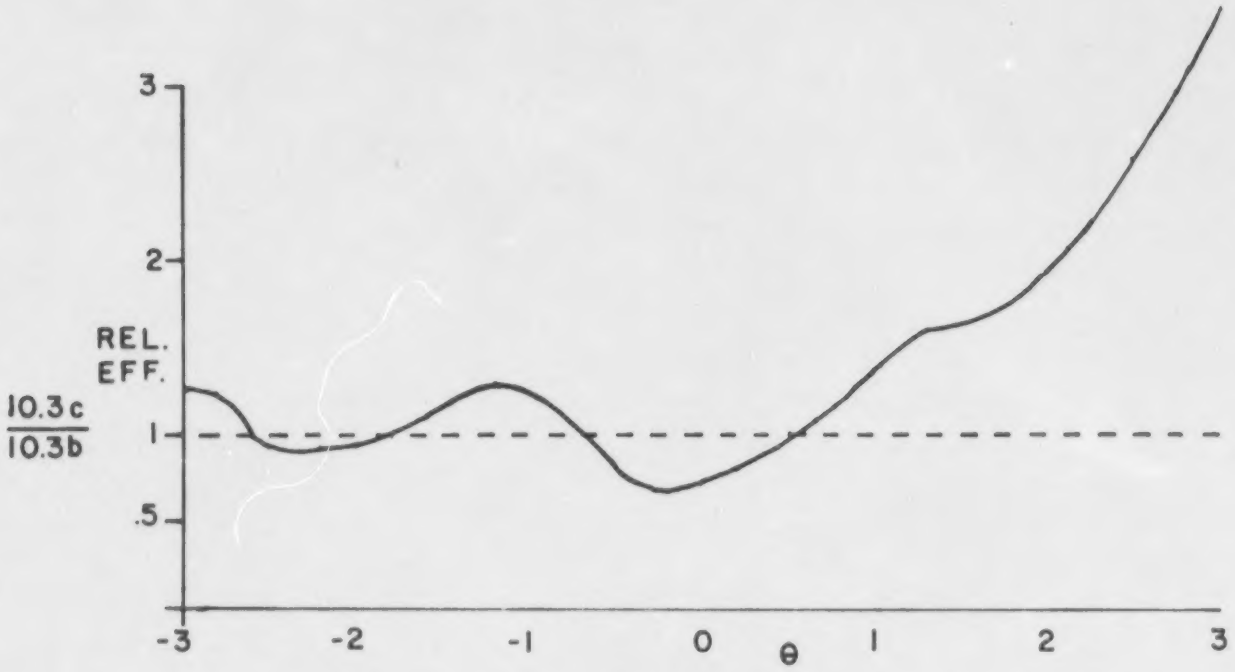


Figure 10.4. The Relative Efficiency Curve comparing Test Information Curve in Figure 10.3c to that in Figure 10.3b.

Sometimes a test is designed for more than one purpose, such as to be used with two cut scores for entrance into two different schools. In this case a two-humped TIC will give good information at the two cut scores. (e.g. see Figure 10.3c).

A TIC of any desired shape may be constructed, provided the items with the necessary IIFs are available to construct the TIC.

10.4 Usually we already have a test and want to revise it to make it better serve our purpose. A comparision of the new and old versions should be made using the Relative EfficiencyCurve (REC). The REC is nothing more than the ratio of the TICs. The ratio of the two curves is found by dividing the $I(\theta)$ of one test by the $I(\theta)$ of the other test at each point on θ. Figure 10.4 is the REC, comparing the TIC in Figure 10.3c to the TIC in Figure 10.3b.

Where the REC is above 1.0, the test in Figure 10.3c(the test for which the $I(\theta)$ is the numerator of the REC ratio) is better than the test for Figure 10.3b. Where the REC is below 1.0, the test for Figure 10.3b is better. And where the REC = 1.0, the two tests are the same.

By starting with an old test, making substitutions of items, and calculating the REC, you can experiment with and improve the old test by trial and error. It does not take long to develop some skill in replacing items to improve the TIC as desired.

10.5 Every test has some error in it. The Standard Error of Estimate (S.E.E.) is the expected standard deviation of errors of estimated ability. That is, if we were to give a test to a group of examinees with identical θs, and estimate their θs with the test, the standard deviation of those estimates would be the S.E.E.

10.6 If the estimate of θ is a maximum likelihood estimate (see Chapter 12), the S.E.E. at a particular θ is easy to calculate from the TIC. The S.E.E. is equal to the square root of the reciprocal of the height of the TIC (I (θ)).

$$SEE = \frac{1}{\sqrt{I(\theta)}}$$

Since I(θ) varies along the θ scale, so will the S.E.E. The larger I(θ) is, the smaller the S.E.E. A small S.E.E. at a cut point is highly desirable.

10.7 The average S.E.E. $\overline{(S.E.E.)}$ over examinees is related to the reliability of Classical Test Theory (r_{xx}), when the scores are standardized to a standard deviation = 1.0."

$$r_{xx} = 1 - \overline{S.E.E.}^2$$

This relation implies that a test with high reliability may be a poor test for your purposes because it has low information at the critical values of θ. Similarly, a test with low reliability may be an excellent test for some purposes, if it has high information where it is needed. Thus, reliability is highly misleading as to the value of a test.

The relation also makes clear the dependence of reliability on the distribution of ability. If many examinees are on the θ scale where there is high information, then the reliability will be higher than if they are distributed on θ at points where information is low.

CHAPTER 11
The Score Information Curve

11.1 The test information curve ($I(\theta)$) gives the maximum amount of information about θ that can be extracted from the test. However, to get the maximum information, items must be optimally weighed. The optimal weight ($W(\theta)$) of an item is given by

$$W_i(\theta) = \frac{P_i'}{P_i Q_i} = \frac{1.7a e^{1.7a(\theta-b)}}{c + e^{1.7a(\theta-b)}}$$

There is a curious characteristic of $W(\theta)$. It varies with θ. That means that item A should receive different weights for examinees with different θs. But to get $W(\theta)$, you must know θ, which is what you are trying to get by giving the test.

11.2 There are two ways to approach this dilemma.

(1) The most satisfactory way is to use an iterative computer program, such as LOGIST or OGIVIA (see Chap. 15). These computer programs, in effect, make use of the optimal item weights and hence yield maximum information about θ.

(2) A rough approximation would be to take raw scores on the test, divide the distribution of raw scores into, say, top, middle and bottom groups and then rescore using different item weights for each group. This procedure would not yield maximum information, but would provide more information than not using variable item weights at all.

11.3 If neither of the options in Section 11.2 is possible, then you may have to resort to the use of number-right score. In this case the amount of information provided by this scoring procedure becomes of interest. The amount of information provided by a number-right score is called the number-right Score Information Curve (SIC). The formula for the SIC (also written as $I(\theta,X)$) is

$$I(\theta,x)=\frac{(\Sigma P\hat{i})^2}{\Sigma P_i Q_i}$$

11.4 The SIC usually has the same general shape as the TIC, but is lower than the TIC at all values of θ. At high θ the TIC and SIC will be nearly the same height (i.e. SIC/TIC $\approx$ 1.0). As θ becomes smaller and smaller, SIC/TIC becomes smaller. This result means that, at high θs little information is lost by using a number-right score, but at low θs relatively much information is lost. Such is the penalty for use of the inefficient number-right score.

11.5 The SICs of two tests may be used just as the TICs are used. A rough approximation of the standard error of estimate may be found for each θ using the number-right scoring procedure, and the ratio of the SICs of two number-right scored tests may be interpreted in the same manner as the Relative Efficiency Curve for TICs. (Strictly speaking, for this interpretation to be legitimate, the test score must be shown to be an unbiased estimate of θ.)

11.6 The SIC is plotted by a computer program available from the Educational Testing Service (See Chapter 15), and may be derived from a program by John Gugel (see Section 15.4).

CHAPTER 12

Maximum Likelihood Estimation of θ

12.1 There are two main ways in IRT to estimate an examinee's θ. They are called the Maximum Likelihood Estimation method and the Bayesian Modal Estimation method. Both methods use the actual response pattern of the examinee rather than the raw score. The difference between the two methods is merely an additional assumption made by the Bayesian method.

12.2 A response is indicated by the lower case letter u. If the examinee gets item i correct, then $u_i=1$, and if he gets it wrong, then $u_i=0$. A response pattern is also called a response vector, and is represented by the uppercase letter U. A response pattern is a list of zeroes and ones, indicating which questions the examinee got correct or wrong in the order the items appear in the test. For example; in a four-item test, an examinee who got the first two items correct and the last two wrong would have a response pattern U = 1100. If he got the first and third items correct and the other two items wrong, his response pattern would be U = 1010. If he got the first three wrong and the last item correct, he would have a response pattern U = 0001.

12.3 We recall that $P_i(\theta)$ is the probability that an examinee with ability θ will get item i correct. $Q_i(\theta)$ is the probability that an examinee with ability θ will get item i wrong. $Q_i(\theta)=1-P_i(\theta)$. We will abbreviate $P_i(\theta)$ and $Q_i(\theta)$ by P_i and Q_i.

12.4 Probability theory tells us that the probability of independent events occurring together is equal to the product of their separate probabilities. We know that the probability of getting one item correct or wrong is independent of the probability of getting other items correct or wrong for any given value of θ. We know this because of the assumption of local independence.*

12.5 Therefore, the probability of an examinee getting item 1 correct and item 2 wrong is P_1Q_2. The probability of getting both items wrong is Q_1Q_2. Getting item 1 correct and item 2 wrong is the response pattern U=10. Therefore, $P(U=10)=P_1Q_2$, $P(U=00)=Q_1Q_2$, $P(U=01)=Q_1P_2$, and $P(U=11)=P_1P_2$.

Similarly, for three items for a given θ, if:

$P_1 = .3$ $\quad Q_1 = .7$
$P_2 = .6$ $\quad Q_2 = .4$
$P_3 = .8$ $\quad Q_3 = .2$

*The assumption of local independence will be discussed in Sec. 14.3.

then

U	L(U|θ) = Likelihood	$\pi_1^3 P_i^u Q_i^{1-u}$
000	$Q_1Q_2Q_3$ = .7 x .4 x .2 = .056	
001	$Q_1Q_2P_3$ = .7 x .4 x .8 = .224	
010	$Q_1P_2Q_3$ = .7 x .6 x .2 = .084	
100	$P_1Q_2Q_3$ = .3 x .4 x .2 = .024	
011	$Q_1P_2P_3$ = .7 x .6 x .8 = .336	
101	$P_1Q_2P_3$ = .3 x .4 x .8 = .096	
110	$P_1P_2Q_3$ = .3 x .6 x .2 = .036	
111	$P_1P_2P_3$ = .3 x .6 x .8 = .144	

Table 12.5

The likelihood of each possible response pattern for a given θ where the $P_i(\theta)$ is as given in Section 12.5.

12.6 These probabilities are called likelihoods (and written $L(U|\theta)$).

Each likelihood is the conditional probability of a response pattern (U) given θ, i.e. $L(U|\theta)$. The general formula for a likelihood is

$$L(U|\theta) = \pi_{i=1}^{n} P_i^u Q_i^{1-u}$$

The upper case Greek letter $\prod_{i=1}^{n}$ means the product of all the $P_i^{u}Q_i^{1-u}$'s where i goes from 1 to n (n = the # of items in the test), just as, in statistical notation $\sum_{i=1}^{n}$ means the sum of a series of numbers where i goes from 1 to n.

When $u_i = 1$

$$P_i^{u}Q_i^{1-u} = P_i^{1}Q_i^{1-1} = P_i^{1}Q_i^{0} = P_i^{1}\cdot 1 = P_i$$

When $u_i = 0$

$$P_i^{u}Q_i^{1-u} = P_i^{0}Q_i^{1-0} = P_i^{0}Q_i^{1} = 1\cdot Q = Q_i$$

When $u_i = 1$, the Q_i drops out, and when $u_i = 0$, the P_i drops out.

Thus, $P^{u}Q^{1-u}$ is just a convenient mathematical way of getting rid of the P or Q depending on the value of u_i. For a three-item test the likelihood of U = 011,

$$L(U=011\,|\,\theta) = \prod_{i=1}^{3} P_i^{u}Q_i^{1-u} =$$

$$= P_1^{u}Q_1^{1-u}\cdot P_2^{u}Q_2^{1-u}\cdot P_3^{u}Q_3^{1-u} = P_1^{0}Q_1^{1-0}\cdot P_2^{1}Q_2^{1-1}\cdot P_3^{1}Q_3^{1-1} =$$

$$= P_1^{0}Q_1^{1}\cdot P_2^{1}Q_2^{0}\cdot P_3^{1}Q_3^{0} = Q_1\cdot P_2\cdot P_3$$

	#1		#2		#3			
θ	P_1	Q_1	P_2	Q_2	P_3	Q_3	L(U=010|θ)	L(θ|U)
-3.0	.29	.71	.36	.64	.21	.79	.71 x .36 x .79 = .202	.169
-2.5	.32	.68	.39	.61	.22	.78	.68 x .39 x .78 = .207	.173
-2.0	.37	.63	.45	.55	.25	.75	.63 x .45 x .75 = .213	.178
-1.5	.50	.50	.60	.40	.30	.70	.50 x .60 x .70 = .210	.176
-1.0	.62	.38	.77	.23	.38	.62	.38 x .77 x .62 = .181	.151
-0.5	.77	.23	.90	.10	.50	.50	.23 x .90 x .50 = .104	.087
0.0	.88	.12	.97	.03	.59	.41	.12 x .97 x .41 = .048	.040
0.5	.93	.07	.99	.01	.70	.30	.07 x .99 x .30 = .021	.018
1.0	.97	.03	.99	.01	.79	.21	.03 x .99 x .21 = .006	.001
1.5	.98	.02	.99	.01	.87	.13	.02 x .99 x .13 = .003	.000
2.0	.99	.01	.99	.01	.91	.09	.01 x .99 x .09 = .000	.000
2.5	.99	.01	.99	.01	.95	.05	.01 x .99 x .05 = .000	.000
							$\Sigma L(U\|\theta) = 1.195$	1.000

Table 12.7

The method of calculating the Maximum Likelihood Estimate of θ from a test of 3 items for an examinee with the response pattern, U = 010.

12.7 When we give a test, we get each examinee's response pattern, and we want his θ. $L(U|\theta)$ is not what we want, since we already have U. What would help us estimate an examinee's θ is just the reverse, i.e. $L(\theta|U)$.

Fortunately, Bayes' Theorem allows us to get $L(\theta|U)$ from $L(U|\theta)$.

$$L(\theta|U) = \frac{L(U|\theta)}{\Sigma L(U|\theta)}$$

To use Bayes' Theorem we have to get the $L(U|\theta)$ at several points on the θ scale. How many points we use is determined by how accurately we want to estimate θ.

To show how this is done, $L(U=010|\theta)$ is calculated in Table 12.7 for three hypothetical items at 12 values of θ.

The total of the $L(U|\theta)$'s is $\Sigma L(U|\theta)$. The right column shows $L(\theta|U) = L(U|\theta)/\Sigma L(U|\theta)$. Any examinee, no matter what his θ, could conceivably have a U = 010 in this three-item test. There is a finite probability of U = 010 at every θ.

However, the likelihood of an examinee having U = 010 varies considerably with θ. An examinee with $\theta \geq 0.0$ is unlikely to have U = 010. In fact, only 6% of examinees with $\theta \geq 0.0$ will have U = 010.

Note: The proponents of Maximum Likelihood Estimation do not agree with the use of Bayes' Theorem in this explanation.

A graph of the likelihoods (for U = 010) would look like Figure 12.7

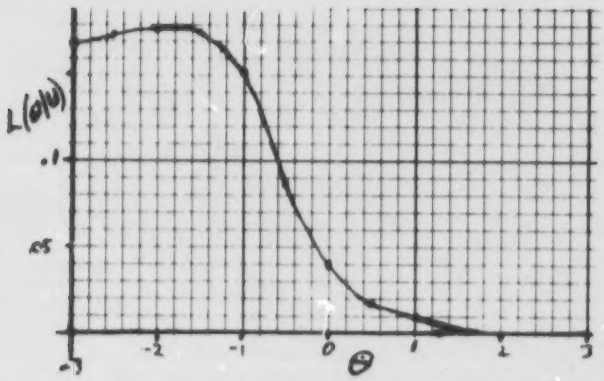


Figure 12.7. The graph of the likelihoods in Table 12.7, called the likelihood function.

This curve is called the likelihood function.

If you had to guess the θ of an examinee with U = 010, what θ would you guess from the information in Table 12.7? You should guess his θ = -2.0 because the likelihood of U = 010 is greater at θ = -2.0 than at any other θ. Therefore, you would be right more often than if you guessed any other θ. By choosing the θ with the greatest likelihood, you have chosen the θ with the maximum likelihood. And that is the Maximum Likelihood method of estimating θ! That's all there is to it.

Now look at the $L(U|\theta)$ column. At which value of θ is $L(U|\theta)$ greatest? It is at $\theta = -2.0$, the same as the θ with the maximum $L(\theta|U)$. That will always be the case because the $L(\theta|U)$'s are just the $L(U|\theta)$'s divided by the constant $\Sigma L(U|\theta)$. So the θ with the maximum $L(\theta|U)$ will always be the same as the θ with the maximum $L(U|\theta)$. Therefore, it is not necessary to divide by $\Sigma L(U|\theta)$ in order to find the θ with the maximum likelihood.

Since we divided by $\Sigma L(U|\theta)$ in order to apply Bayes' Theorem, we find that Bayes' Theorem is not necessary for maximum likelihood estimation.

Another short cut is to take the logarithm of the P_i and Q_i's and add them, instead of multiplying the P_i's and Q_i's. The sum of the logarithms will also always be maximum at the same value of θ. A graph of the log likelihoods is called the log likelihood function. The log likelihood function will always be highest at the same θ at which the likelihood function is highest.

It should be noted that, in this example, you would be right in estimating $\theta = -2.0$ only 17.8% of the time and wrong 82.2% of the time. But this is true only because the test had only three items. With a longer test there would be one θ at which the likelihood is much greater than any other.

12.8 Table 12.8 shows the maximum likelihood method of estimating θ for a test made of the four items whose IRF's are shown in Figure 6.17.

(1) across the top are 17 values of θ

(2) under the θ's are the $P(\theta)$'s for each of the four items.

(3) the item numbers and parameters are in the top left corner.

(4) down the left side are the 16 possible response patterns for four items and the raw (# right) score represented by the response patterns.

a	b	c	WO-8 CGK Item #		-2.7	-2.3	-2.0	-1.7	-1.3	-1.0	-.7	-.3	0.0	.3	.7	1.0	1.3	1.7	2.0	2.3	2.7	
3.76	1.80	.00	17		.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.04	.35	.78	.96	.99	
1.00	-.96	.26	21	P(θ)	.30	.33	.37	.43	.53	.62	.71	.82	.88	.92	.96	.97	.98	.99	.99	.99	.9	
1.46	-1.24	.36	47		.38	.40	.45	.52	.66	.77	.87	.94	.97	.99	.99	.99	.99	.99	.99	.99	.99	
.62	-.05	.15	50		.20	.23	.25	.28	.33	.38	.44	.52	.59	.65	.74	.79	.84	.89	.91	.94	.96	
P(θ)= True Score=					.88	.96	1.07	1.23	1.62	1.77	2.02	2.28	2.44	2.56	2.69	2.76	2.85	3.22	3.67	3.88	3.93	
Raw Score	Response Pattern 17 21 47 50																					MLE
0	0	0	0	0	347	310	200	197	107	54	21	5	0	0	0	0	0	0	0	0	0	-00
1	0	0	0	1	87	92	87	77	53	33	17	6	2	1	0	0	0	0	0	0	0	-2.3
	0	0	1	0	203	206	212	213	208	181	141	81	48	28	10	1	0	0	0	0	0	-1.7
	0	1	0	0	149	152	153	149	121	88	52	24	11	3	2	2	2	1	0	0	0	-2.0
	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.001	.004	.007	.006	.004	+2.0
2	0	0	1	1	53	62	75	83	102	111	110	88	69	29	29	23	16	6	2	0	0	-1.0
	0	1	0	1	37	46	51	58	59	54	41	26	16	6	7	8	8	6	2	0	0	-1.3
	0	1	1	0	91	102	125	161	234	296	346	370	350	319	247	200	149	70	19	2	0	-.3
	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	.002	.007	.030	.071	.090	.095	>2.7
	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	.062	.127	.381	.695	.570	.392	+2.0
	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	.021	.063	.381	.695	.570	.392	+2.0
3	0	1	1	1	23	30	42	63	115	181	272	401	504	592	703	751	782	567	196	37	9	+1.3
	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	.29	.67	3.08	7.03	8.93	9.40	>2.7
	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	.077	.329	3.08	7.03	8.93	9.41	>2.7
	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2.0	6.2	37.7	68.8	56.5	38.8	+2.0
4	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	8	33	305	696	884	931	+00
	Total				1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	.00	

Table 12.8

An illustration of the MLE of θ for all possible response patterns from a test composed of four real items. (All likelihoods are multiplied by 1000 to reduce decimal values).

(5) in the body of the table are the $L(U|\theta)$'s for each possible U for the 17 values of θ. Each $L(U|\theta)$ is multiplied by 1000 to eliminate decimal values.
(6) underlined in each row is the maximum $L(U|\theta)$
(7) down the right side are the values of θ where the underlined maximum likelihoods occur. These θ's are the maximum likehood estimates (MLE) of θ for each of the 16 possible U.

Note that the MLE for U = 0000 is $-\infty$, and the MLE for U = 1111 is $+\infty$. That is a characteristic of the MLE. The MLE will not give a finite estimate of θ unless the examinee has missed at least one item and answered at least one item correctly. This limitation is not serious because raw scores of 0% or 100% are usually rare.

The MLE of $\theta > 2.7$ is due to the limited range of θ used in this example. A larger range of θ would yield a more precise MLE of θ.

The many cells with $L(U|\theta) = 0$ in the body of Table 12.8 are due to the very unusual item #17.

12.9 Now compare in Table 12.8 the raw scores on the left with the MLE's on the right. You can see that a raw score of 1 represents θs from -2.3 to +2.0, an extreme range! A raw score of 2 represents θs from -1.3 to greater than +2.7. A raw score of 3 represents θ's from +1.3 to greater than +2.7.

The extreme range of θ, depending on the U's represented by a single raw score, demonstrates well the inadequacy of using raw score as an estimate of ability. The inadequacy of raw score as an estimate of ability is due to the fact that raw score cannot distinguish chance success from knowledge success on an item. In contrast, the MLE takes guessing into account by using the additional information in the response pattern.

CHAPTER 13

Bayesian Modal Estimation of θ

13.1 The Bayesian Modal method of estimating θ takes up where the MLE stops. The proponents of the Bayesian Modal method (called Bayesians) reason that if the distribution of θ is known or assumed, then that knowledge or assumption provides additional information which can be used to more accurately estimate θ.

13.2 Bayesians assume that θ is distributed normally. The assumption of normality means that the probability of any randomly-chosen examinee having a θ at the extremes is less than his probability of having a θ located near the mean. The assumption of normality is made on an a priori basis (i.e. before empirical evidence). Thus, it is called the normal "prior" distribution.

13.3 Suppose the likelihood of $\theta_1|U$ is very close to the likelihood of $\theta_2|U$, but that there are many more examinee's at θ_2 than at θ_1. In this case we would be right more often by estimating θ at θ_2 than at θ_1. In doing so we would, in effect, be weighting our likelihood by the number of examinees at the two θ values. If we take this idea to its logical extreme, we should weight all likelihoods by the proportion of examinees at each value of θ in order to reduce our errors.

13.4 By assuming a normal distribution of θ we can weight the likelihood by the relative proportions of area under the normal curve. To do this we merely multiply the area within the interval of the normal curve at θ, designated $\int N(0,1)$, times $L(U|\theta)$. Table 13.4 shows how this is done

θ = Ability/Knowledge Scale

	Midpoint				-2.7	-2.3	-2.0	-1.7	-1.3	-1.0	-.7	-.3	0	.3	.7	1.0	1.3	1.7	2.0	2.3	2.7	
	Upper Limit				-2.50	-2.15	1.85	-1.50	-1.15	-.85	-.50	-.15	.15	.50	.85	1.15	1.50	1.85	2.15	2.50	2.85	
	Lower Limit				-2.85	-2.50	-2.15	-1.85	-1.50	-1.15	-.85	-.50	-.15	.15	.50	.85	1.15	1.50	1.85	2.15	2.50	
Raw Score	Response Pattern 17	21	47	50	Area .004	.0094	.0164	.0346	.0583	.0726	.1108	.2427	.1192	.2427	.1108	.0726	.0583	.0346	.0164	.0094	.004	B.M.E.
0	0	0	0	0	38.8	91.4	426	682	624	392	233	121	119	0	0	0	0	0	0	0	0	-1.7
1	0	0	0	1	34.8	86.5	143	266	309	240	188	66	24	24	0	0	0	0	0	0	0	-1.3
	0	0	1	0	81.2	194	349	737	1212	1314	1560	897	572	680	110	7	0	0	0	0	0	-.7
	0	1	0	0	60	143	251	516	705	639	576	266	131	73	22	15	12	3	0	0	0	-1.3
	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.0075	.0133	.0702	.0060	0	2.0
2	0	0	1	1	21	58	123	287	595	806	1230	975	822	704	321	167	93	21	0	0	0	-.7
	0	1	0	1	15	43	84	201	344	392	454	288	191	146	78	58	47	21	0	0	0	-.7
	0	1	1	0	36	96	205	557	1360	2150	3830	4100	4170	7740	2740	1450	870	240	0	0	0	+.3
	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	.017	.039	.108	.116	.870	0	+2.3
	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	.45	.74	1.32	1.14	5.50	0	+2.3
	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	.15	.37	1.32	1.14	5.50	0	+2.3
3	0	1	1	1	9	28	69	218	670	1314	3010	4440	6010	14400	7790	5450	4560	1960	300	0	0	+.3
	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2	4	10.6	11.5	9	0	+2.0
	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	.6	1.9	10.7	11.5	8.6	0	+2.0
	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	15	36	130	113	540	0	+2.3
4	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	58	192	1055	1140	8000	0	+2.3

Table 13.4

An illustration of the Bayesian Modal Estimate of θ for all possible response patterns from a test composed of four real items. (All likelihoods are multiplied by 10,000 to reduce decimal values).

using the likelihoods from Table 12.8.

(1) the top row are points of θ which are midpoints of intervals of θ.

(2) the 2nd and 3rd rows are the limits of the intervals.

(3) the 4th row is the proportion of area under the normal curve and within the interval.

(4) in the body of the table each column is the area in the 4th row multiplied by the corresponding likelihood from Table 12.8 (times 100,000 to remove decimal values, i.e., $L(U|\theta) \times \int N(0,1)$).

(5) the largest value in each row is underlined.

(6) the θ for the underlined likelihoods are in the right column. These are the Bayesian Modal Estimates (BME) of θ.

The BME is called modal because, when we choose the largest value in each row, we are choosing the mode of the distribution of $L(U|\theta) \times \int N(0,1)$.

13.5 Bayesian Modal Estimates are more conservative than MLEs (conservative means closer to zero, the mean of the normal prior distribution). Note that with U=0000 and U=1111, the BMEs of θ are finite. The finiteness of θ estimates of BME when either all or no items are answered correctly is a minor advantage of BME.

"*Note: There are several computational errors in Table 13.4. Howev these errors do not affect the explanation of the concepts involved

13.6 There is an active controversy between the Bayesians and the proponents of the MLE. The Bayesians argue that MLE is the same as a BME, if θ is assumed to be distributed rectangularly. (A rectangular distribution of θ means that there are equal numbers of examinees at all θ values, even at $+\infty$ and $-\infty$). And so, say the Bayesians, since a normal distribution of θ is more reasonable to assume than a rectangular distribution, the BME is a more accurate estimate of θ.

The proponents of MLE argue that the coincidence of the MLE (which assumes no distribution of θ) being the same as a BME with rectangular distribution is irrelevant. The important thing is that MLE makes no assumption about the distribution of θ, whereas BME makes the additional assumption, which will be sometimes false.*

13.7 I shall not take sides in this matter, because for me the point is moot. The only computer program available to me at present is OGIVIA-3 (See Chap. 15), which uses the BME. Therefore, I shall continue to use BME until I have a program which uses MLE. At that time I shall have to make a decision.

13.8 Another type of Bayesian estimation is called Owen's Bayesian, after its inventor, R. L. Owen (1975). The Owen's Bayesian method is used primarily in tailored testing (See Chap. 17).

*I apologize to both sides of this complex issue for this meager representation of their positions.

CHAPTER 14
Assumptions

14.1 There are 4 basic assumptions of IRT. The first of these is a minor assumption. It is an assumption of any test theory and without which there would be no justification for testing.

Assumption #1: The Know-Correct Assumption: if the examinee knows the correct answer to the item, he will answer it correctly.* We have probably all violated this assumption while taking tests by marking a different choice than we intended to mark. Occasionally, an examinee will inadvertently skip an item, and then mark all the rest of his answers in the wrong places. This is merely a clerical error, but there is no provision for it in any test theory. Another way to state the first assumption is: if he got the item wrong, then he did not know the answer.

14.2 Assumption #2: The Normal Ogive Assumption: The IRF takes the form of the normal ogive. This is the problem, mentioned in Section 3.3, which deterred Lord's work for 10 years. The difficulty lay with 3 parts of the IRF.

a. The lower asymptote
b. The upper asymptote
c. The middle or rapidly rising part of the IRF

*The reader should take careful note that the inverse of this assumption is NOT made. That is, it is NOT ASSUMED that if the examinee gets the item correct, he knows the answer. I emphasize this distinction because many persons upon first reading of assumption #1 misread it as its inverse.

(1) As previously noted, the c-value of an IRF is often not 1/A. This is the case with observed parts of the lower asymptote. But what about the unobserved parts? If an item from the SAT with c = .09 were given to extremely low θ persons such as kindergarten children or mentally retarded persons, would the lower tail of the IRF rise to 1/A?

(2) It has been charged by Hoffman (1962), that tests may penalize extremely high ability persons, because they know too much. That is, they consider factors far beyond the intended scope of the item, and therefore get it wrong. If that were the case, then the IRF would curve down away from the upper asymptote at high θ's. This has been called the Banesh Hoffmann Effect.

(3) It was not known that the IRF was monotonic, and that its general shape was that of a normal ogive.

In 1965 Lord published a massive study with a sample size greater than 100,000. Specifically, he found:

a. the lower tail of the IRF did not rise for almost all items. The very few items that did rise, did so to a very small extent.

b. no evidence of the Banesh Hoffman Effect.

c. good indications that the IRF is strictly monotonic.

14.3 Assumption #3: Local Independence. Local independence means that the probability of an examinee getting an item correct is unaffected by the answers given to other items in the test. Local independence does NOT mean that the items correlate zero with each other.

The most common situation where local independence does not hold is in a speeded test. In a speeded test an examinee may get the last items wrong, simply because he did not reach them. A distinction is made between not-reached items and omitted items. Not-reached items are those unanswered items which have no answered items after them in the sequence of items in the test. Omitted items (omits) are unanswered items which have at least one item answered after them in the sequence of items in the test. This distinction is important, when deciding what to do with not-reached items and omits in scoring answer sheets. Not-reached items are not attempted (and hence there is no possibility of being correctly answered) simply because of the presence of the early items, which were attempted during the time limit.

Furthermore, earlier items which were attempted may have been missed, because the examinees felt rushed and could not give their best efforts to the items.

Similarly, in long tests, fatigue effects may impact the local independence of items.

Certain reading comprehension tests might violate local independence when several items are all based upon some common reading passage. However, it is not entirely clear whether such items violate local independence.

Chain items violate local independence. An example of chain items follows:

(1) Who discovered America?
(2) Where was he born?

Clearly, if the first item were not in the test, the second item would be meaningless. Fortunately, chain items are rare.

Local independence also means that items are uncorrelated for individuals with the same θ. This interpretation suggests a statistical test for local independence. (Lord, in preparation, p. 26).

$$r_{gh} / \theta = 0, \; g \neq h$$

where r_{gh}/θ = the tetrachoric correlation between items g and h for examinees with exactly the same ability.

To use this statistical test requires that first it is necessary to get a large number of examinees with identical θ's. Then, using their responses, calculate the interitem tetrachoric correlation. That correlation should not be significantly different from zero. This procedure has at least 2 practical difficulties.

First, it should be done for all (or at least several) values of θ. It is nearly impossible to get large sample sizes at many θ values.

Second, it must be done for all pairs of items, which would require calculation of $n(n-1)/2$ tetrachoric correlations (n = # of items in the test) for each value of θ. A 50-item test would require 1225 correlations at each θ value. If 10 θ values were chosen, that would mean 12,250 correlations.

A similar but simpler procedure would be to partial out of the interitem correlations the affect of θ. This may be done by using the item-test biserial correlation. Then

$$r_{gh \cdot \theta} = \frac{r_{gh} - r_{g\theta}\, r_{h\theta}}{\sqrt{1 - r^2_{g\theta}}\ \sqrt{1 - r^2_{h\theta}}}$$

where r_{gh} = the tetrachoric correlation among all examinees between items g and h ($g \neq h$), and $r_{g\theta}$ = the biserial correlation between item g and the θ. $r_{gh|\theta}$ should not be significantly different from zero. Before using this test of local independence, care should be taken that the implicit assumptions of the statistics involved are satisfactorily met. In any case it should only be considered as a rough estimate.

This latter procedure would require n(n-1)/2 tetrachoric correlations plus n biserial correlations (which are usually available anyway).

Because of the practical rarity of conditions violating local independence, this assumption is usually not tested.

14.4 Assumption #4: Unidimensionality. The assumption of unidimensionality is the most complex and most restrictive assumption of IRT. In general, unidimensionality means that the items measure one and only one area of knowledge or ability. However, unidimensionality does NOT mean that the items correlate positively with each other. In fact, it is conceivable for all items to correlate negatively with each other and still be unidimensional.

As a rule of thumb, tests that look unidimensional probably are unidimensional. Thus, typical ability tests, such as verbal, numerical, spatial perception, mechanical comprehension and tool knowledge are probably unidimensional.

Another rule of thumb is, items that test bits of knowledge that were learned together are probably unidimensional. Thus, a final examination for a college course might be considered unidimensional. An excellent example of this rule is given by Bejar, Weiss, and Kingsbury (1977). That study involved a test in college introductory biology. Part of the course was covered by a test divided into 3 content areas, called "Chemistry," "The Cell," and "Energy." The single test for all 3 content areas was found to be essentially unidimensional.

Unidimensionality in a test covering 3 such diverse sounding content areas is surprising. The fact of its unidimensionality may have resulted from the items testing bits of knowledge which were learned together in the college course.

It may well have been, however, that the subject-matters of the 3 content areas were not as diverse as they sound. It is likely that "Chemistry" was the chemistry necessary to understand the cell. And "The Cell" content was necessary to understand the "Energy" use and transfer within the cell. This possibility suggests another rule of thumb. Items that test bits of knowledge which are logically and sequentially related may be expected to be unidimensional.

Rules of thumb are, by definition, sometimes erroneous. I do not suggest that they replace efforts to empirically verify unidimensionality. However, in view of the difficulty of empirical verification, some readers may find them helpful.

14.5 There is no completely satisfactory test for unidimensionality among multiple-choice items. The reason for this situation is that most tests for unidimensionality involve factor analysis of interitem tetrachoric correlations. Unfortunately, the tetrachoric correlation assumes θ is normally distributed, and is not entirely appropriate when $c \neq 0$; i.e., when the item can be correctly answered by guessing. Cristofferson (1975) has made the best attempt to develop a test for unidimensionality (Lord, in preparation, Section 2.4, p.27). However, the mathematics of this method are complex and will not be discussed.

I have found 8 methods of testing for unidimensionality in the literature. Six of the eight use factor analysis. To avoid repetition, the initial factor analysis steps which are common to all six will be described.

(1) convert the actual responses of examinees into zeroes and ones; zero, if the response is wrong, and one, if the response is correct. Factor analysis requires a sample 10 times the # of items ($N = 10n$);

(2) calculate a matrix of interitem tetrachoric correlations (not the phi coefficient), using the zero-one responses;

(3) replace each value in the diagonal with the correlation in its row that has the largest absolute value (most factor analysis computer programs have an option to do this automatically). If there are too many items for the capacity of the computer, a random sample of items may be used;

(4) do a principal component (or principal axis) factor analysis for the first 9 factors (9 is an arbitrary, typical number).

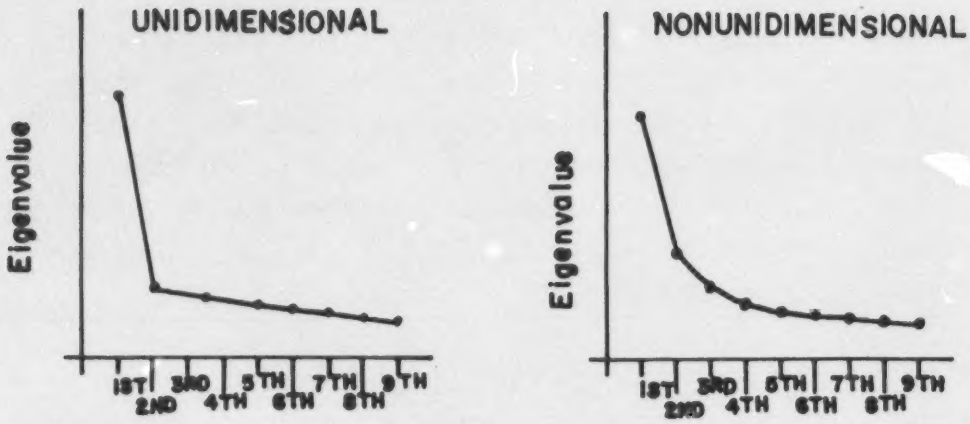
14.6 I have given short titles for easy reference to each of the tests for unidimensionality:

(1) The Eigenvalue Test. Plot the eigenvalues of the nine factors against the factor rank, as shown in Figures 14.6(1)a and 14.6(1)b. The items may be considered unidimensional if the eigenvalue of the first factor is large compared to the second factor, and the eigenvalues of the remaining factors are all about the same. The graph should look something like Figure 14.6(1)a if the items are unidimensional, and like Figure 14.6(1)b if the items are not unidimensional. (Lord and Novick, 1968, p.283).

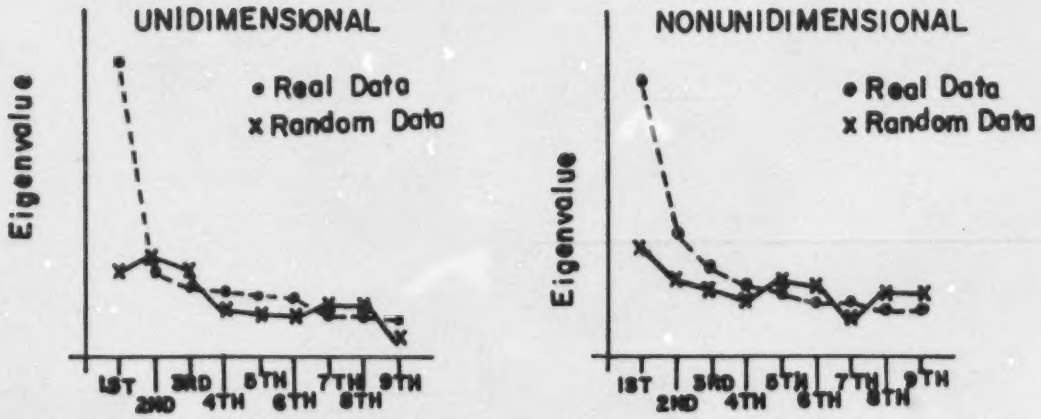
(2) The Random Baseline Test. This test is a variation of the Eigenvalue Test. It is necessary to do the Eigenvalue Test first. To get the random baseline, create with a random generator a matrix of zeroes and ones of the same order as the matrix in step (1) of Section 14.5. Then perform steps (2), (3), and (4) just as with the Eigenvalue Test. Plot the eigenvalues from the random data on the same graph as the Eigenvalue Test. Unidimensionality is indicated if only the first factor eigenvalues are distinguishable for the 2 sets of data (McBride and Weiss, 1974,p.30). See Figures 14.6(2)a and 14.6(2)b.

(3) The Biserial Test. Compute the correlation between the item-test biserial correlation and the item first factor loading. A high (.80 or higher) correlation supports the assumption of unidimensionality. (McBride and Weiss, 1974,p.31,33 and 37).

(4) The Factor Loading Test. Unidimensionality is indicated if the first factor loadings for all items are significant and have the same sign (+ or -). (McBride and Weiss, 1974,p.33).



Figures 14.6(1) a and b. A hypothetical illustration of the Eigenvalue Test for unidimensionality.



Figures 14.6(2) a and b. A hypothetical illustration of the Random Baseline Test for unidimensionality.

(5) The Congruence Test. If the examinees can be separated into two meaningful subgroups (Black/White, male/female), then a factor analysis (steps (1) to (4) in Section 14.5) can be performed on each group. The Coefficient of Congruence (C_{AB}) of the item first factor loadings between the 2 groups will approach one, giving evidence of unidimensionality with respect to the variable on which the groups were defined, i.e. race, sex. (Pine, 1977,p.4). (See Rommell, 1970,p.461 for Coefficient of Congruence).

$$C_{A,B} = \frac{\sum(L_{iA} \cdot L_{iB})}{\sqrt{(\sum L_{iA}^2)(\sum L_{iB}^2)}}$$

L_{iA} = loading of item i for group A on the 1st factor
L_{iB} = loading of item i for group B on the 1st factor
n = number of items in the test
$C_{A,B}$ = Coefficient of Congruence between groups A & B

(6) The Communality Test. It has been suggested that unidimensionality may be tested by

$$G = \frac{\sum \frac{r_{ij}}{\sqrt{h_i^2 \; h_i^2}}}{n}, i > j$$

where r_{ij} = interitem tetrachoric correlation

h_i^2 = the item communality

n = the number of correlations

According to Green, et al (1977, p.836) this function which I have designated G, approaches 1.00 as dimensionality approaches unity.

I have applied G to data published in McBride and Weiss (1974). This data gives item communalities and interitem tetrachoric correlations for six real, word knowledge tests, and six sets of random data. The six real tests were found by McBride & Weiss to be essentially unidimensional by three different measures of unidimensionality, i.e. the Random Baseline Test (see 14.6(2) above), the Biserial Test (see 14.6(3) above), and the Factor Loading Test (see 14.6(4) above). G for the real tests ranged from .419 to .484, and had a Spearman rank correlation with the first factor percent of common variance of rho = 1.00. On the random data, G ranged from .284 to .348, and had a Spearman rank correlation with the first factor percent of common variance of rho = .60. It appears that when $c \neq 0$, G approaches neither one (for unidimensionality) nor zero (for nonunidimensionality). Furthermore, G is no better as an indicator of unidimensionality than is the first factor percent of common variance.

(7) The Part/Whole Test. If the items may be separated into distinctive types or content, the a and b values may be estimated separately for each type and for the entire test. If the parameter estimates under the two conditions (part vs. whole) correlate highly, unidimensionality is supported (Bejar, 1977(b), p.13).

(8) The Vector Frequency Test. Assuming θ is normally distributed, and given the item parameters, it is possible to calculate the expected frequency of all possible response patterns. A comparison with the observed frequency of all possible response patterns will yield a <u>non-significant</u> chi-square, if unidimensionality is present (Bock and Lieberman, 1970).

14.7 Unidimensionality is a sufficient condition for local independence. That is, if you have unidimensionality, then you also always have local independence. The reverse is not true. Local independence is necessary for, but does not guarantee, unidimensionality.

CHAPTER 15

Computer Programs

15.1 There are several computer programs available for estimating examinees' θs and item parameters. Only 2 of those are in general use. Both are written in FORTRAN.

15.2 LOGIST was written at the Educational Testing Service and is the program used by Lord for his work (See Wood et al, 1976). The LOGIST and related programs provide a complete set of options for calculating and printing:

a. examinee's θ,
b. item parameters,
c. item response curves,
d. test characteristic curve,
e. item information function,
f. score information curve, and
g. relative efficiency of 2 tests.

LOGIST allows either examinee's θ or item parameters as fixed input, and puts all other estimates on the same scale as the input parameters. It is by far the more versatile program. Lord recommends that, to get good estimates, at least 1000 examinees and 30 items are needed in the test.

However, LOGIST has one practical disadvantage. It requires from 30 minutes to two hours of computer CPU (Central Processing Unit) time. Consequently, I was unable to convince my data processing people to implement LOGIST and have not been able to use it.

LOGIST uses a maximum likelihood estimation procedure. It computes all parameters at the same time, using an iterative technique. The iterative technique computes the first estimates from the raw data. Then, those estimates become input for the second iteration of computation, using the same maximum likelihood procedure to compute the second estimate. The second estimate becomes input for the third, and so on. The iterations continue until the estimates converge, and do not change significantly from one iteration to the next. Sometimes the estimates do not converge, but drift off to infinity or fluctuate wildly back and forth. In these cases, LOGIST applies certain limiting rules.

The a and b parameters from LOGIST correlate positively. This is an unexpected and undesirable result. When c parameters do not converge, LOGIST sets all non-converging c parameters equal to some average value, usually between .10 and .25. This may occur with 50% to 80% of the items in a single test, which suggests that the c parameter is not well estimated by LOGIST.

15.3 OGIVIA* was written for Dr. Vern Urry of the U.S. Civil Service Commission (USCSC) by Jerry Edwards, University of Washington and revised by John Gugel of the USCSC. It has also been called URRY and ESTEM in the literature. There are several versions of it, the current one being called OGIVIA-3. OGIVIA-3 calculates and prints both a classical item analysis and the item parameters. It has options for the normal ogive and logistic models, but does not have the scaling option of LOGIST. It does not print out examinees' θ's, but could be made to do so without much trouble.

*Pronounced ogive-eye-aye

OGIVIA uses a Bayesian modal estimation procedure. It estimates item parameters, using raw scores as an estimate of θ, by fitting the data to a logistic (or normal) ogive. Chi Square is the test for goodness of fit. It then re-estimates θ with the estimated item parameters using an iterative technique until the θ estimates converge or 20 iterations are done, whichever comes first. The re-estimates of θ are then used to re-estimate the item parameters by the same curve fitting technique. Estimates of θ, typically, do not converge on a small percent (about 1%) of examinees; and item parameters sometimes do not converge on as many as 5% to 10% of the items.

OGIVIA needs at least 1000 examinees and 60 items in the test, with a test KR-20 of +.90 in order to get good estimates. I have used it with as few as 150 examinees on a test of 30 items with apparent success for my purposes. Uses of such small numbers should be done with caution. OGIVIA requires only two to five minutes of CPU time on the computer. This fact makes OGIVIA much more attractive than LOGIST, despite the possible difficulties of Bayesian estimations.

An interesting feature is an F-ratio for each item, which tells how well the items responses fit the model. An F-ratio of 4.00 or 5.00 or less means the data fit the model. In a comparison of the F-ratios of 8 tests between the normal ogive model and the logistic model, the logistic model fit the data slightly better than the normal ogive model.

Urry has a new version of OGIVIA, called ANCILLES, which needs only 30 items, but little is known about it because it is so new.

The a and b parameters of OGIVIA do not correlate highly. However, the c parameter estimates fluctuate considerably from sample to sample. This indicates that OGIVIA, too, does not estimate the c parameter well.

15.4 Programs which print out IRFs, IIFs and TICs are available from the USCSC. These were written by John Gugel.

15.5 All of these computer programs are available for the asking.

CHAPTER 16
Equating the θ-Scales

16.1 When raw data is fed into LOGIST or OGIVIA, the program calculates the item parameters (called "calibrating" the item) and the examinees' θs all at once. The a and b values are on the same scale as θ, which is set to have a mean = 0, and standard deviation = 1. If the same test is given to two groups of examinees, and the items calibrated separately for each group, the a and b values from the separate calibrations will not be comparable because the scales of θ will be on different metrics. And the examinees θs will not be comparable from group to group. All that is necessary to correct this situation is to lump both groups together and treat them as one group. Then the θs for all examinees will be on the same scale.

16.2 Another (more laborious) method is to transform the θ-scale for one group to the θ-scale of the other group. This transformation is possible because the b-value of an item is invariant (except for linear transformations of the θ-scale), and because the θ-scales are linearly related.

The linear relationship is based on the traditional standardization formula

$$\frac{b_2 - \bar{b}_2}{SD_{b_2}} = \frac{b_1 - \bar{b}_1}{SD_{b_1}} \qquad \text{Eq. 16.2a}$$

where b_1 = the item b-value on the metric of Group 1,

b_2 = the item b-value on the metric of Group 2,

$\bar{b}_1$ & SD_{b1} = the mean and standard deviation of the b-values of the items on metric of Group 1,

$\bar{b}_2$ & SD_{b_2} = the mean and standard deviation of the b-values of the items on metric of Group 2.

Lord (in preparation) recommends that, when calculating the $\bar{b}$ and SD_b, items with low a values ($a < .8$) and extreme b-values ($|b| > 2.0$) be excluded, because b-values for such items are not well estimated.

Solving equation 16.2a for b_1,

$$b_1 = \left[\frac{SD_{b_1}}{SD_{b_2}}\right] b_2 + \left[\bar{b}_1 - \left(\frac{SD_{b_1}}{SD_{b_2}}\right)\bar{b}_2\right] \qquad \text{Eq. 16.2b}$$

Since b is on the θ-scale, θ may be substituted for b, and

$$\theta_1 = \left[\frac{SD_{b_1}}{SD_{b_2}}\right] \theta_2 + \left[\bar{b}_1 - \left(\frac{SD_{b_1}}{SD_{b_2}}\right)\bar{b}_2\right] \qquad \text{Eq. 16.2c}$$

This equation may be used to transform an examinee's θ score from one scale to another. It is NOT proper to use a regression equation based on the correlation of the b-values or θ from the two groups of examinees for this purpose.

16.3 Suppose from a bank of 100 items we construct two tests containing the following items:

Test	Item Bank #s
A	1-60
B	41-100

Each test has 60 items, 20 of which are common to both tests. Suppose further, we give the tests to two groups of examinees (one test to each group), and calibrate the tests separately.

We now want to put all 100 items on the same scale. We can do so because we have 20 common items on the two tests and we have those items calibrated on both scales. First, calculate the mean and standard deviation of the b-values on each scale of the 20 common items. Then all the b-values of one test may be converted to the metric of the other test with Equation 16.2b. The θs may be converted with Equation 16.2c.

The a-values are converted by dividing by the ratio of the b-value standard deviations.

$$a_1 = a_2 \div \frac{SD_{b1}}{SD_{b2}} = a_2 \cdot \frac{SD_{b2}}{SD_{b1}}$$

NOTE: Remember, a_2 and b_2 means the a and b values on the old scale and a_1 and b_1 are the item's a and b values on the new scale.

The c-values are already on the same scale, because they are on the P(θ) axis of the IRF. Thus, all c-values are always on the same scale and need no conversion.

In order to build a large bank of calibrated items on the same scale, it is desirable to include in the test items from the bank which have already been calibrated along with new items. These items, which are used to link one θ-scale to another θ-scale, are called "anchor items." A minimum of 17 anchor items is recommended. More than 17 is desirable.

16.4 Occasionally the situation will arise where two different tests (i.e., no common items) are given to two groups of examinees at different times, and some of the examinees take both tests (called "anchor persons"). This situation may be handled in either of two ways.

(1) If there are enough persons, combine the answer sheets of the anchor persons for both tests, and treat the two tests as one long test. Then the θ-scales of the two separate tests may be rescaled to the combined test θ-scale as described in Section 16.2 above.

(2) Calibrate each test separately. Take the two θ values of the anchor persons and calculate their means and standard deviations for each test. Then use:

$$\theta_1 = \left[\frac{SD_{\theta_1}}{SD_{\theta_2}}\right] \theta_2 + \left[\bar{\theta}_1 - \left(\frac{SD_{\theta_1}}{SD_{\theta_2}}\right) \bar{\theta}_2\right]$$

to rescale all examinee's θs,

$$b_1 = \left[\frac{SD_{\theta_1}}{SD_{\theta_2}}\right] b_2 + \left[\bar{\theta}_1 - \left(\frac{SD_{\theta_1}}{SD_{\theta_2}}\right) \bar{\theta}_2\right]$$

to rescale the b-values, and to rescale a-values:

$$a_1 = a_2 \cdot \left(\frac{SD_{\theta_2}}{SD_{\theta_1}}\right)$$

Again, the c-values are already on the same scale.

The use of the θs to rescale assumes that θ has not changed between administrations of the two tests.

16.5 Even if no anchor persons have taken both tests, it still may be possible to take items from both tests, create a third test, administer it to a third group of examinees, and then use the third test as anchor items to link the original 2 tests.

16.6 Anchor items should be chosen in order to make the estimate of SD_b as accurate as possible. All estimates of b-values have some error in them. To reduce the proportionate contribution of estimation error to SD_b we want the SD_b as large as possible. That conclusion suggests that the anchor item b-values should have a bimodal distribution. That is, half of the anchor items should have high b-values, and half should have low b-values.

However, very high and very low b-values are not estimated well, which means they have a significant amount of error in them. We should, therefore, compromise between a large SD_b and large error in estimates of b. This reasoning suggests that anchor items should be bimodally distributed with half the items having moderately high b-values, say $1 \leq b \leq 1.5$, and half with moderately low b-values, i.e., $-1.5 \leq b \leq -1.0$.

The b-values are with respect to the groups to be tested. If there is good reason to believe that the group to be tested has about the same distribution of ability as those on whom the anchor items were calibrated, then the bimodally distributed b-values are the best anchor items. However, if the group turns out not to have about the same distribution of ability as the calibration group, half the anchor items may be either too easy or too hard, and the anchor items will not serve their purpose well.

A safer method would be to select anchor items to have a rectangular distribution of b-values from -1.5 to 1.5. In this way you will be confident of getting many anchor items of appropriate difficulty and still have a large SD_b.

The a-values of anchor items should be as large as possible, and the c-values as small as possible. However, the a-value is more important than the c-value.

CHAPTER 17
Tailored Testing

17.1 Psychometrists long have known and deplored the fact that many items on a test are not appropriate for a given examinee, i.e. they are either too hard or too easy. Until IRT there was no satisfactory way to avoid this problem, and at the same time get a decent measure along the ability scale.

With IRT came the possibility of tailored testing, which is so named because it allows the "tailoring" of the test to the ability of the examinee. Tailored testing is also called adaptive testing. Variations of it are called stradaptive testing and flexilevel testing.

17.2 Tailored tests are administered by a computer with the items presented on a CRT (Cathode Ray Tube device, which is similar to a television set). (See Ree, 1977a.) It works like this:

(1) The examinee sits in front of a CRT attached to a typewriter keyboard.

(2) The examinee registers on the computer with his identification, test name and other pertinent information.

(3) In the computer are stored a bank of 150 to 200, or more, precalibrated items along with their item parameters. The computer selects an item of average difficulty and presents the item to the examinee on the CRT.

(4) The examinee records his answer on the typewriter keyboard.

(5) The computer uses the examinee's response and the item parameters to estimate the examinee's most likely θ, and then selects another item. The item selected is the one which will best help the computer estimate θ after the examinee answers the item. If the examinee got the item correct, he will get a different next item than if he got the item wrong.

(6) Steps (4) and (5) above are repeated until the computer meets the criterion for stopping the test. The criterion for stopping the test is called the "stopping rule."

17.3 Examinees with different response patterns will, in general, get a different set of items; yet their final estimates will be on the same metric. Not all examinees may get the same number of items, yet all θ estimates can be to the same degree of accuracy.

17.4 Stopping rules can be designed as desired to fit the situation. Three typical stopping rules are:

(1) Stop when a specified number of items have been administered.

(2) Stop when the SEE of the examinee's θ has dropped below a specified value; often SEE $\leq .0625$ is used.

(3) Stop when no more items remain in the bank that will provide a significant amount of information about the examinee.

It is not uncommon to combine some of the above rules.

17.5 Because the computer selects items on the basis of item information, the computer will usually select items with high a-value first, and then after high a-value items have been exhausted, select other items.

17.6 The reader will recall from Section 12.8 that the maximum likelihood method (MLE) will estimate θ at plus or minus infinity until the examinee gets one item wrong and one item correct. Therefore, if the examinee gets the first item correct, the computer will give the hardest item in the bank second. And it will continue to give the hardest items until the examinee gets one wrong.

Similarly, if the examinee gets the first item wrong, the computer will give the easiest items until the examinee gets one correct. Since Bayesian estimation methods (see Chapter 13) do not have this characteristic, it has been proposed to combine the two estimation methods, using Bayesian estimation until the examinee has gotten one item wrong and one item correct; and then switch to MLE.

Owen (1975) has developed a highly efficient algorithm for Bayesian scoring. The Owen's Bayesian scoring procedure is widely used.

17.7 Tailored testing has several advantages over conventional tests.

(1) Depending upon the characteristics of the item bank, a tailored test will use only 10% to 50% of the number of items required by a conventional test and at the same time will measure more accurately than the conventional test at almost all values of θ. Tailored tests can measure to any specified degree of accuracy.

(2) A tailored test takes much less time to administer, or several abilities can be measured by a tailored test in the same time needed to measure one ability by a conventional test.

(3) Security of the items is much improved, because different examinees get different items, and because the items are much less accessible (in the computer as opposed to hard copy).

17.8 The use of tailored testing also has some problems.

The cost of large scale use of tailored testing machines is currently prohibitive because of the cost of CRT devices, an on-line time-sharing computer, and telephone lines to hook the CRT devices to the computer. Moreover, it often takes 5 seconds for the computer to do its calculations and present the next item. If only 20 CRT devices were on line at a time, the delay to get the next item could be 100 seconds or almost two minutes. Such delay would wipe out the advantage of reduced administration time that makes tailored testing attractive. The reliability of telephone land lines is also often a problem.

A feasible alternative would be a self-contained tailored testing machine with the items presented on a Microfiche, and the calculations done by a microprocessor.

Major Brian Waters, USAF, has developed a prototype of such a machine, which could be mass-produced for about $500 each. His design requires the examinee to find the item on the Microfiche film. The microprocessor "senses" the location of the film and will not accept a response from the examinee if he is looking at the wrong item. Waters estimates that the microprocessor could be made to control the Microfiche machine (i.e. present the item automatically) for a mass-produced cost of $1500 each.

Another potential problem is that of legal defensibility. Imagine an examinee who, after talking to another examinee, finds out that his items were different, he got a different number of items, he got more items correct, and yet got a lower score on the test. This situation contains all the necessary elements for a law suit. Now imagine trying to explain to a judge or jury that, in fact, the examinee was not improperly discriminated against, and that the Bayesian modal or maximum likelihood method of estimating theta was more accurate. Also consider the hundreds of so-called "testing experts" across the country who have never heard of item response theory and who might be called to testify. You may now have an inkling of the enormous problems ahead for what Lord calls "occult scoring methods."

17.9 Nevertheless, work is progressing toward the use of tailored testing. The U.S. Civil Service Commission has adopted the use of tailored testing as a matter of policy. The U.S. Air Force Human

Resources Laboratory, San Antonio, Texas, has a tailored testing machine operating on an experimental basis at the San Antonio AFEES (Armed Forces Entrance Examination Station). (Ree, 1977a) Several studies of live tailored testing have been published by the Psychometric Methods Program at the University of Minnesota. The Educational Testing Service is also considering tailored testing and intends to engineer its own tailored testing machine.

17.10 Obtaining a large bank of calibrated items is not a simple matter. As a result, the military services have formed an Ad Hoc Group on Adaptive Testing. One of its purposes is to share calibrated items. It has become evident that even the sharing of the items gets complicated. (See Ree, 1978.) Below is a list of information necessary to share items.

(1) item itself with key
(2) reference for the correct answer
(3) a, b, and c values
(4) evidence of goodness-of-fit (i.e. if OGIVIA is used, the Chi-Square and F-ratio).
(5) evidence of unidimensionality
(6) the name of the dimension
(7) the computer program used to estimate parameters
(8) normal ogive or logistic model (on OGIVIA, 1st or 2nd cycle)
(9) description of the sample on whom calibrated, and size
(10) evidence of cultural-fairness
(11) description of anchor items used, if any.

Once the bank has been established, a method for controlling its use must be established. Most users will want the items with high a-values and low c-values. If the users are giving the items to the same population, several users may give the same item to the same examinees. Such over-exposure can destroy the worth of the item.

Items in the bank may be duplicates or near duplicates. Thus, careful visual inspection of the items in the bank will be required.

17.11 Such problems are only a few of those which will be encountered as work progresses. A humorous, actual example can give an idea of how problems cannot be anticipated.

At the San Antonio AFEES, examinees were being tested on a tailored-testing machine. The CRT device was hooked to the computer by telephone line. The telephone used for the connection sat on the table beside the CRT device. On the telephone were two buttons labeled "DATA" and "TALK."

One examinee, when left alone, pressed the "TALK" button, breaking the connection with the computer, and called his mother in Dallas. (Ree, 1977b.)

Murphy's Law reigns supreme.

CHAPTER 18

Item Cultural Bias

18.1 The study of culture-fair testing has become highly complex, since the issue came to public attention in the late 1960's. There are at least 5 statistical definitions of bias and fairness, and 3 ethical positions. This paper will not try to sort out those matters. For some of the more important papers; see Cleary, 1969; Darlington, 1971; Hunter & Schmidt, 1976; and Thorndike, 1971.

Moreover, the issue of the practical effect of test bias on predicting some outside criterion, such as job performance or college GPA is also not of concern here, since we have guaranteed construct validity by the requirement for unidimensionality. In this sense the criterion is the examinee's true θ.

18.2 Studies of item-bias using classical test theory often compare the item p-values for one group with the item p-values of another group. Items with significantly different p-values between the 2 groups are thought to be culturally-biased items. Such an approach is inappropriate for at least 2 reasons. First, the method assumes that the two groups have the same average ability. That assumption is probably false even if the groups are matched on moderator variables such as educational level, since the quality of education varies considerably from school to school.

Second, the comparison of p-values across groups assumes that the bivariate distribution of the p-values is linearly related. Lord (1977) gives several proofs that the p-values can NOT be linearly related. One proof will be described here.

Consider 2 items. Item one is extremely easy, and item two is extremely difficult. Assume both items are administered to two different ethnic groups, A & B, in a test with other items of intermediate difficulty. The p-values for both groups for item 1 will be 1.00 because the item is so easy. The p-values for both groups for item 2 will be the c-value of the item, if the item is so hard that all members of both groups have to guess at the answer. We can plot the points represented by the p-values of these 2 items. (See Fig. 18.2a).

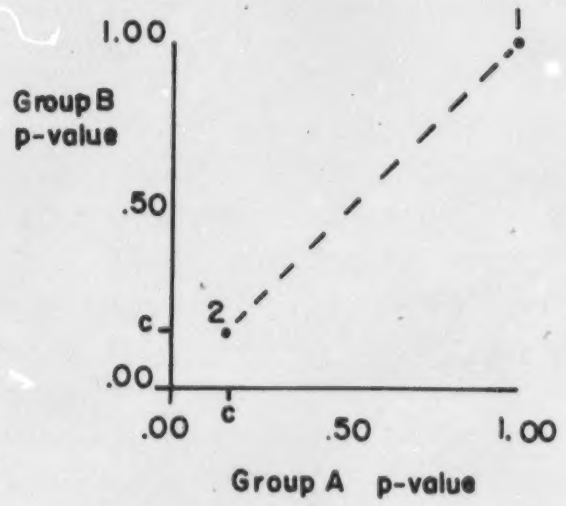


Figure 18.2a. The bivariate distribution of p-values for two hypothetical items, # 1 and # 2, for two hypothetical groups, A and B. Item # 1 is extremely easy, and item # 2 is extremely difficult for both groups.

In order to be linearly related the bivariate distribution of p-values of the other items in the test between groups must fall around the straight dashed line in Fig. 18.2a, connecting the points for items 1 and 2. However, if group A does better as a whole on the test than group B, the p-value points for many items of intermediate difficulty will fall to one side of the straight line. (See Fig. 18.2b).

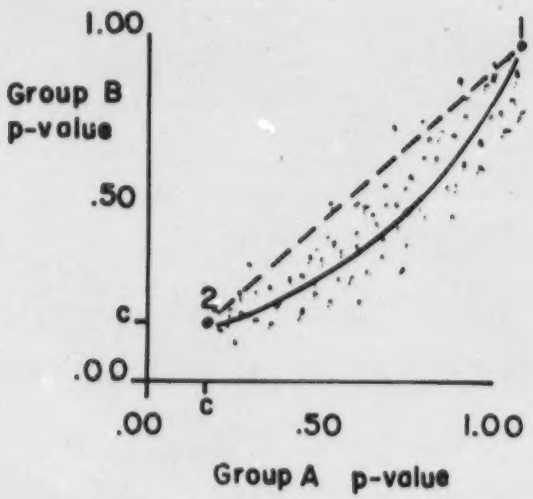


Figure 18.2b. The bivariate distribution of p-values of items in a hypothetical test on which Group A does better than Group B.

The line of best fit then must pass through the points for items 1 and 2, and through the middle of the bivariate distribution of the p-values for other items in the test. That line must be curved to do so, as is the solid line in Fig. 18.2b. Therefore, the relationship between p-values cannot be linear. The same is true of other classical item parameters, such as the "corrected" p-value, the inverse normal transformation of the p-value, and "delta" (Lord and Novick, 1968, p. 381).

For this reason (and others) the use of classical test theory item parameters is inappropriate for, and can lead to erroneous identification of item bias.

18.3 The a,b, and c item parameters are invariant across groups under the assumption of unidimensionality. Therefore, it should not matter in principle on what group the item is calibrated. Whatever the group, the a, b, and c values should be the same except for some random estimate error, once the a and b values are on the same metric.

If the a and/or b and/or c values are significantly different, when calibrated separately on two groups (and put on the same metric), it means that examinees with identical θs will have different chances of getting the item correct ($P(\theta)$), depending on their group. That situation is clearly unfair. Thus, we may define bias between groups A & B as

$$P_A(\theta=k) \neq P_B(\theta=k)$$

where k is some of value of θ.

Of course, if $P_A(\theta) \neq P_B(\theta)$ for one θ-value, then they will not be equal for other θ-values. However, it is not true that they must be unequal for all θ-values. It is quite possible for an item to be biased at, say, high θ and not biased at low θ. This possibility stems from the fact that the $P(\theta)$'s may be different due to any one or more of the 3 item parameters being different.

18.4 If $P_A(\theta) \neq P_B(\theta)$, then the item can be used to distinguish between groups A & B, even in the unusual circumstance of all examinees in both groups having identical θs. This distinction means that the interitem correlation, given θ is not zero.

$$r_{ij|\theta} \neq 0$$

But $r_{ij|\theta} = 0$ is a requirement for local independence (see Sec. 14.3), and local independence is a necessary (but not sufficient) condition for unidimensionality (see Sec. 14.7). Therefore, if $P_A(\theta) \neq P_B(\theta)$, the test is not unidimensional with respect to groups A and B. The problem of item bias, then, is one of violation of the assumptions of local independence and unidimensionality with respect to the groups of examinees. I hereby name this condition "group dimensionality".

18.5 If many of the items are group dimensional, that condition may be detected by the Congruence Test (see Sec. 14.6(5)). If only a few items are biased, the Congruence Test may not be sensitive enough to detect them. In any case we still may wish to know just how a particular item is biased, and what relative effect that bias has on the groups of examinees.

18.6 One method used to make this determination is the comparison of the IRF's of the item for each of the groups. Figures 18.6a to 18.6e show the IRFs of actual items from an experimental form

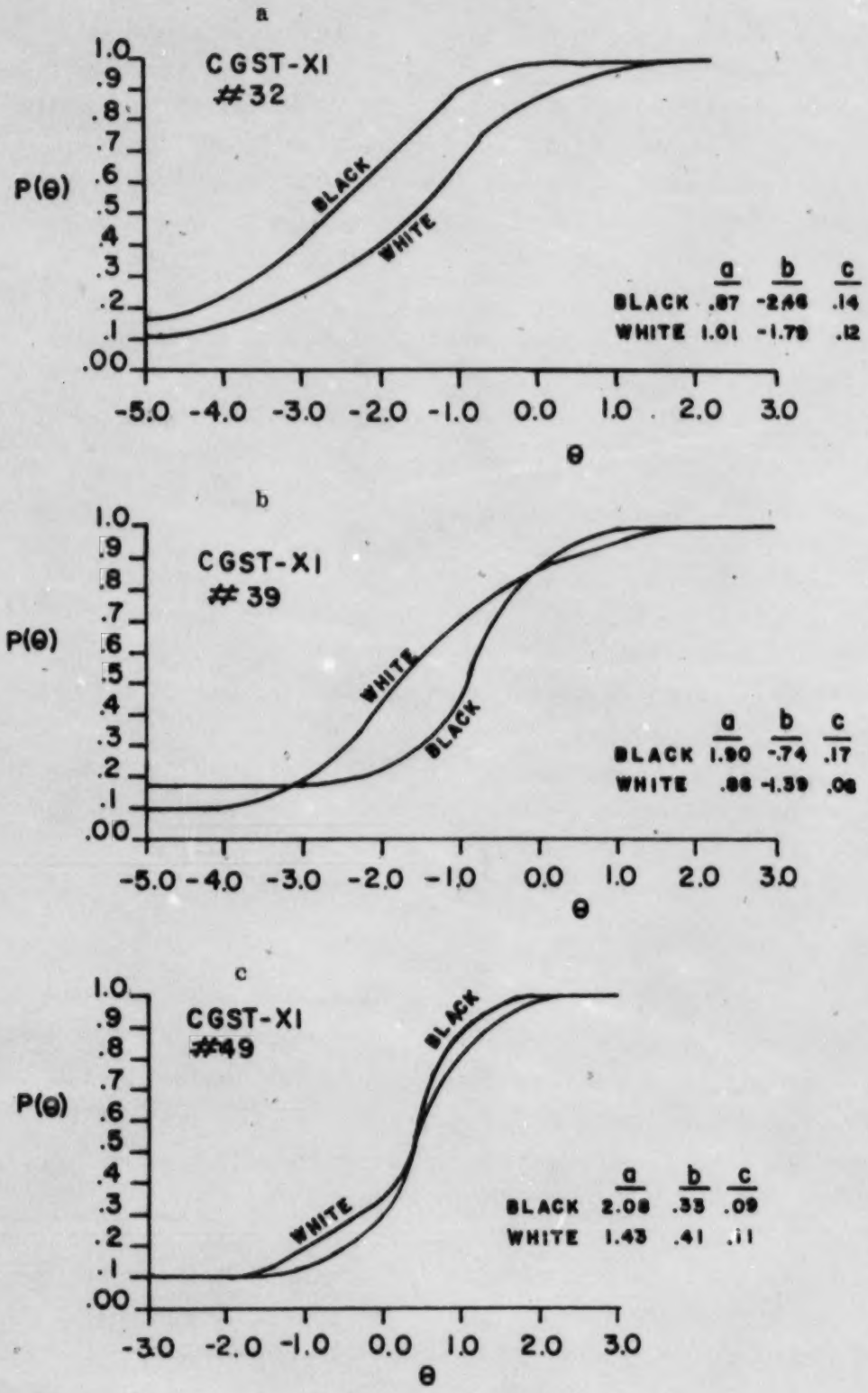
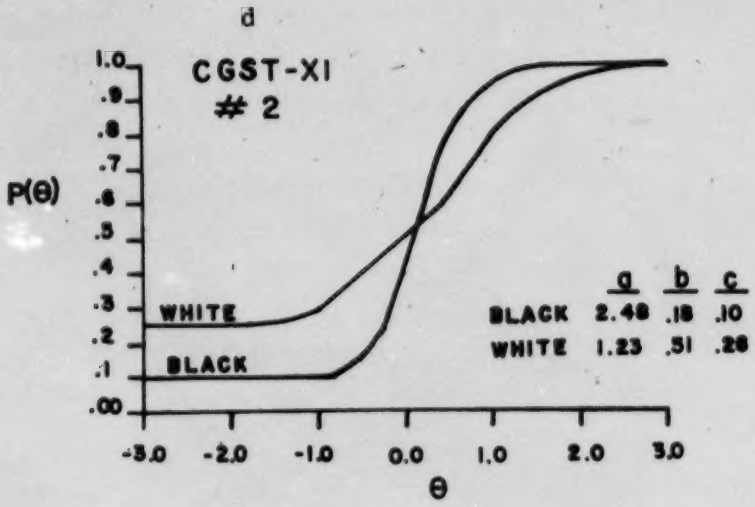
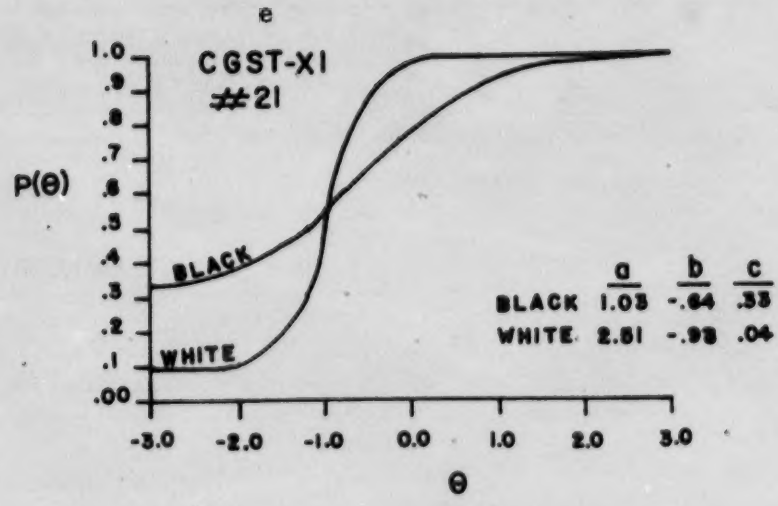


Figure 18.6a, b, and c. Item Response Functions of Blacks and Whites for three real items from an experimental form of the Coast Guard Selection Test.





Figures 18.6d and e. Item Response Functions of Blacks and Whites for two real items from an experimental form of the Coast Guard Selection Test.

of the Coast Guard Selection Test (CGST-X1), when calibrated separately for Blacks and Whites. From Fig. 18.6a we can see that the P(θ) of item #32 is greater for Blacks than Whites at every value of θ, although for θ greater than 0.0 and less than -4.0, they are not much different. Thus, this item is biased in favor of Blacks for θ between -4.0 and 0.0, and fair for both groups elsewhere on the θ-scale. Since the range of θ where this item is biased coincides very closely with the range of θ where the item provides most of its information, the item is not both fair and useful anywhere, and thus is a bad item.

Figure 18.6b shows the IRF's for a similar item, but which is biased in favor of Whites instead. This item is biased for $\theta = -0.6$ to $\theta = -2.4$ and for $\theta < -3.2$. The item is not significantly biased from $\theta = -2.4$ to $\theta = -3.2$ and for $\theta > -0.6$.

Figure 18.6c shows the IRF's of an item that appears fair or nearly fair at all θ in spite of a rather substantial difference in the a-values for the two groups.

The item in Figure 18.6d is biased in favor of Blacks for θ from 0.4 to 2.0, but biased in favor of Whites for all θ less than 0.0.

The item in Figure 18.6e is Black-biased at $\theta < -1.0$, and White-biased for $-0.8 < \theta < 1.0$.

18.7 In my interpretations of the figures in Section 18.6 I have tended to disregard small differences between the IRF's which appear insignificant on the graphs. Actually, it is not yet known how much of a difference between the IRFs of group dimensional items make a significant difference in the estimation of θ. Such a determination would depend upon the distributions of θ for the two groups, which as I indicated in Section 18.2 cannot be expected to be the same.

18.8 In actual practice the item parameters are usually estimated with both groups lumped together in the sample. When this combined-group calibration is done, the resulting item parameters are some complicated (not a simple or even weighted) average of the separate-group calibrated parameters. However, as a rough rule of thumb the IRF of the combined-group falls generally between the IRF's of the separate groups. If the combined-group parameters are then used to estimate θs for both groups, the result will be a non-systematic distortion of the θs for both groups.

18.9 Attempts have been made to develop sample-free indicators of item bias.

One way is to plot the item parameters for the 2 groups as I have done for each of the parameters for Blacks and Whites from CGST-X1 in Figures 18.9a, 18.9b, and 18.9c. The solid lines show the theoretical, expected line of best fit, assuming unidimensionality. The dashed lines were rather arbitrarily chosen to exemplify an acceptance region. If the dashed lines were chosen statistically, items whose item numbers lie outside the region would have statistically, significantly different item parameter(s). Another method is used by Lord (1977). He divided his total sample into 2 random groups, and by conducting separate calibrations on each random group, constructed his own empirical test of significance. After identifying biased items, Lord then repeated the entire 5-step procedure to eliminate the contamination of the biased items. His procedures contemplates the possibility that the presence of strongly biased items may mask the moderate bias of other items.

Rudner (1977) has compared 4 different methods of bias detection, the best of which appears to be the calculation of the area between the two IRFs.

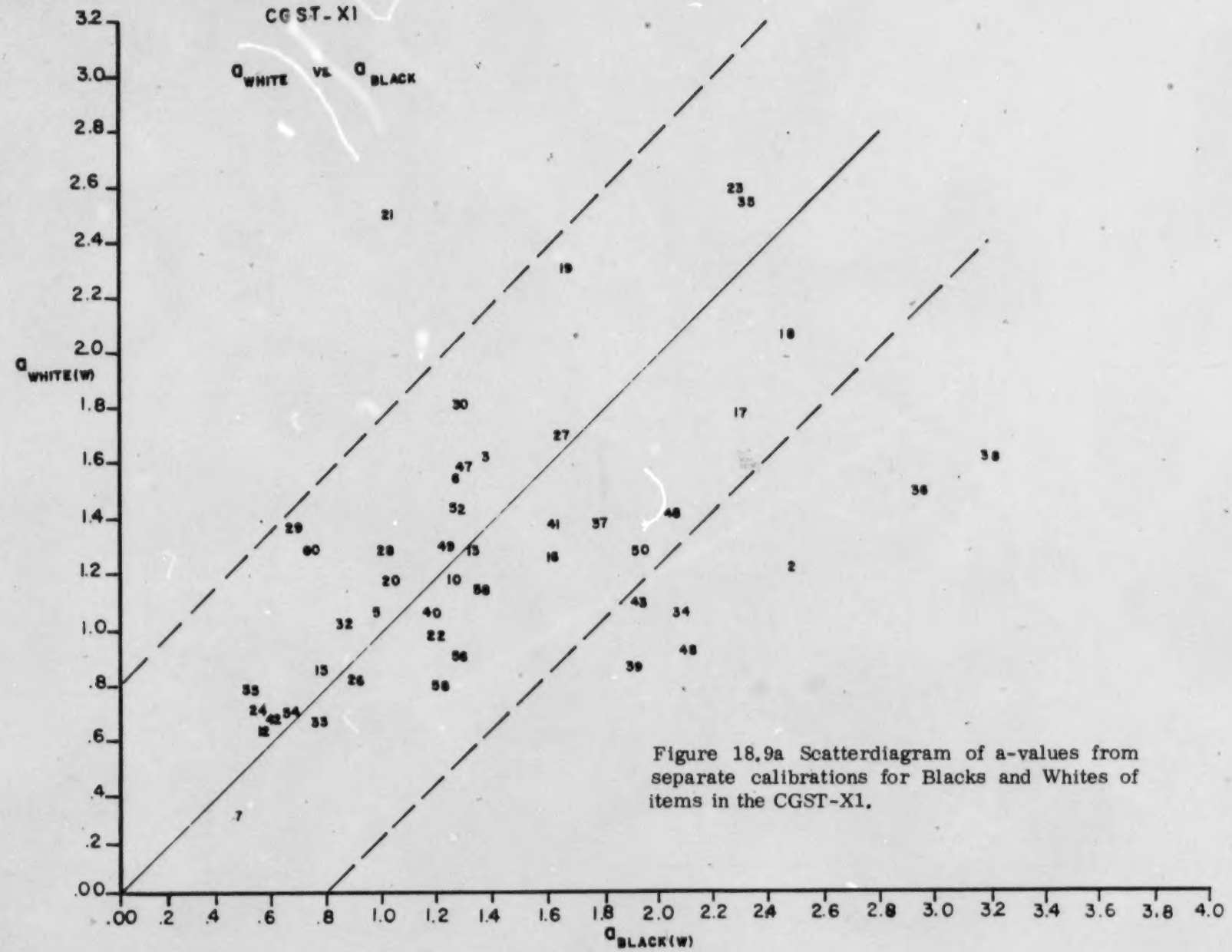


Figure 18.9a Scatterdiagram of a-values from separate calibrations for Blacks and Whites of items in the CGST-X1.

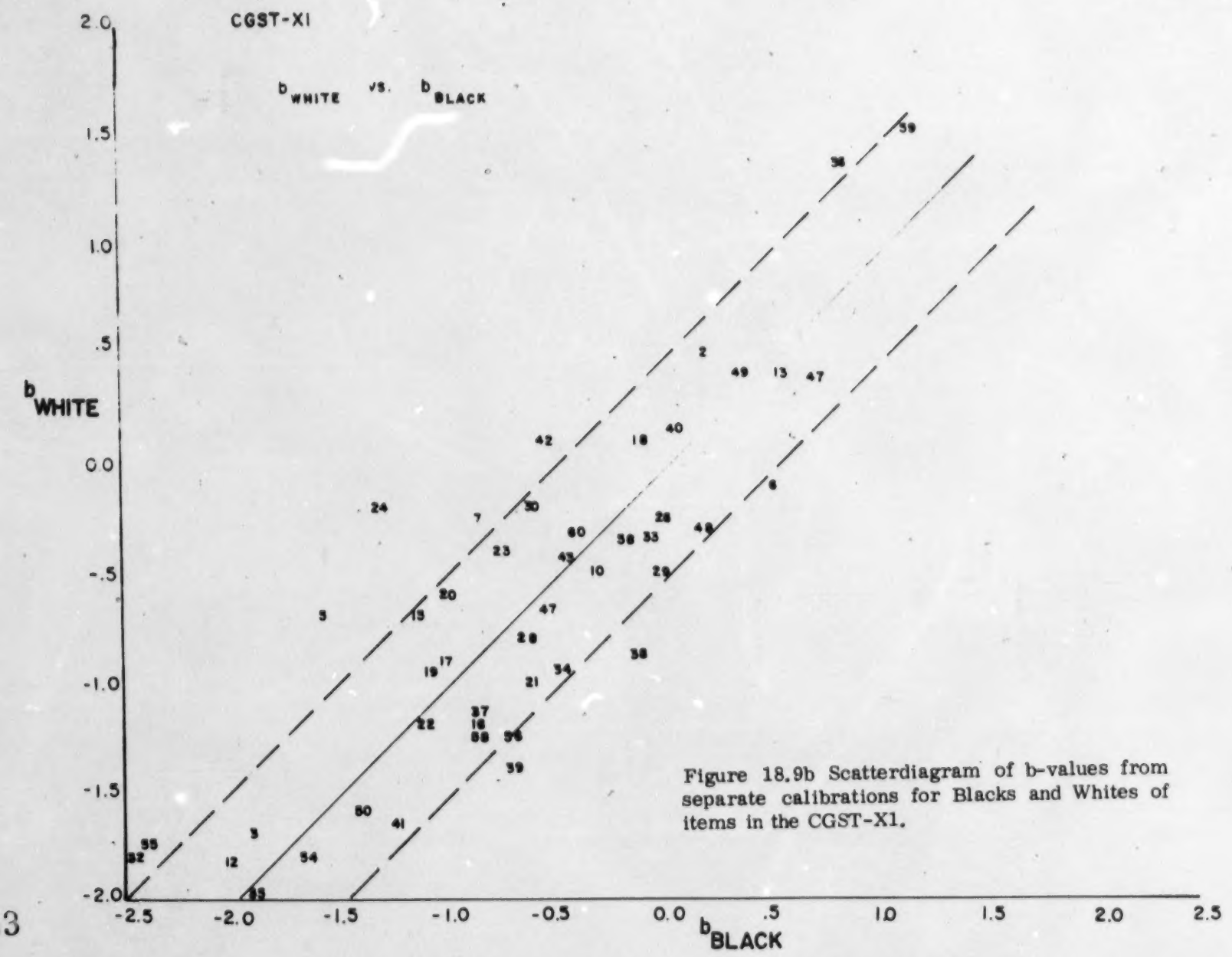


Figure 18.9b Scatterdiagram of b-values from separate calibrations for Blacks and Whites of items in the CGST-X1.

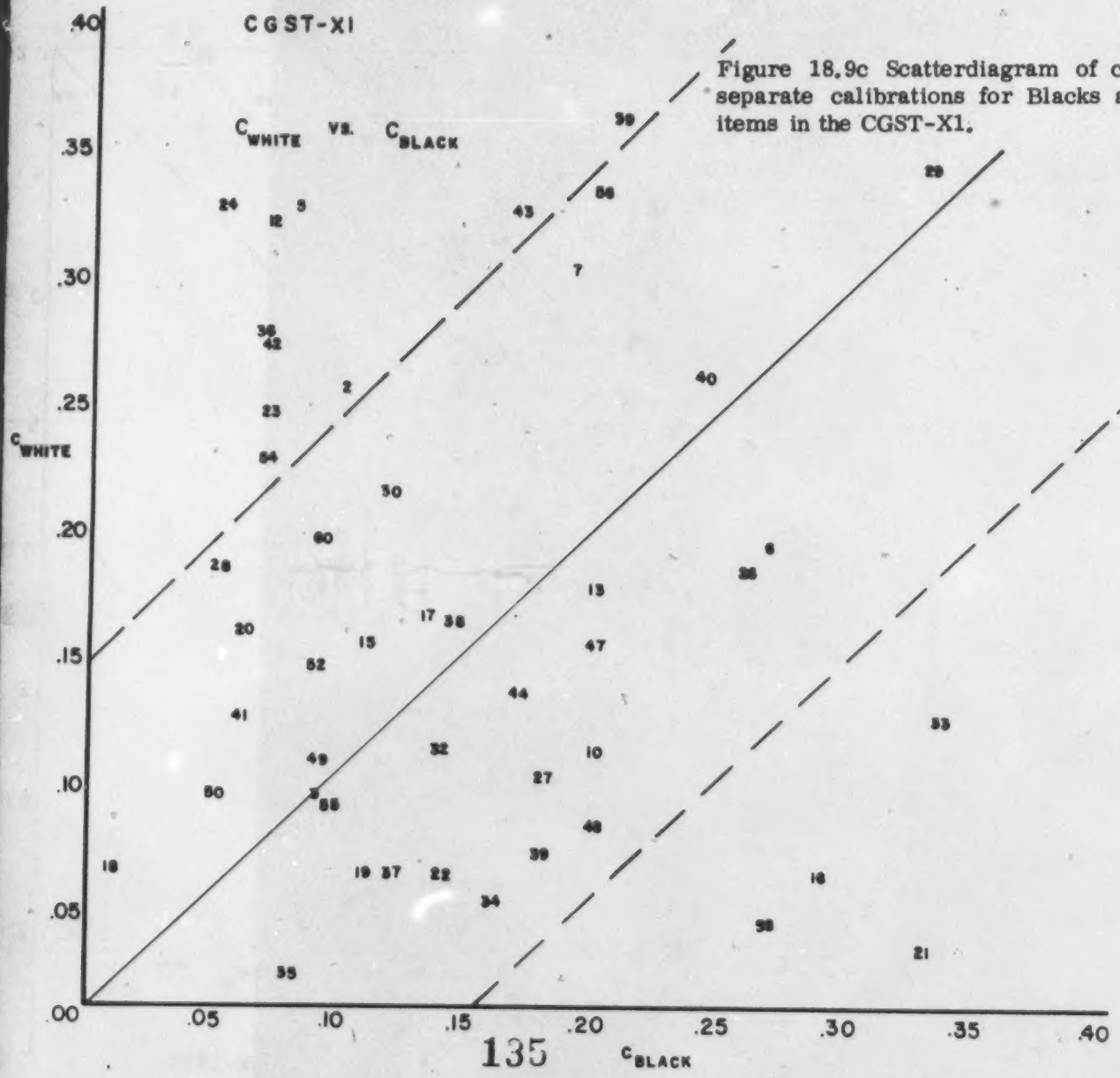


Figure 18.9c Scatterdiagram of c-values from separate calibrations for Blacks and Whites of items in the CGST-X1.

Both methods (Lord's and Rudner's) assume that the c-values for both groups are the same, an assumption which both Lord and Rudner would agree is false.

Lord (in preparation, pp. 296-298) provides true statistical difference tests for maximum likelihood estimates of a-values and b-values (but not c-values).

18.10 What makes an item culturally-biased (i.e., group dimensional)?

In 1968 I conducted research to try to answer this question (using classical test theory). I made two lists of 20 items each. One list contained only Black-biased items, and the other only White-biased items. Even with intense study of the two lists neither I nor Black testing practitioners to whom I showed the lists could come up with any consistent set of hypotheses to explain the bias. My conclusion was that item-bias may not be identified by inspection.

More recent investigations offer some hope. Lord (1977) found a reading comprehension item on Black history in the U.S. to be Black-biased at all values of θ. Durovic (1978) found that two minority reviewers had strong negative reactions to the two items out of 14 that failed the Rasch model test of fit. Scheuneman (1976) found that her chi-square procedure identified items as biased, which contained content readily interpreted as culturally-biased from a common sense point of view.

These results suggest that at least some biased items may be identifiable by inspecting their content. Nevertheless, other item-bias seems inexplicable by content, such as the following two vocabulary items from the Scholastic Aptitude Test. Both items seek the OPPOSITE of the stem word.

2. INJURE (quoted from Lord, 1977, p. 29)
   - A. release
   - B. refrain
   - C. smooth
   - D. embellish
   - * E. heal

8. GEL (quoted from Lord, in preparation, p. 294)
   - A. glaze
   - B. debase
   - C. corrode
   - *D. melt
   - E. infect

Both items are Black-biased (i.e., in favor of Blacks) at low θ and White-biased at high θ.

18.11 Testing practitioners must often make practical decisions even in the total absence of relevant information. Because I am such a testing practitioner, and because of my penchant for rules of thumb, I offer the following guidelines for rejecting items as biased between groups A and B. Let

d(a), d(b), d(c) = the absolute values of the difference of the a, b, and c values, respectively, for the two groups (after being converted to the same scale).

Then, I declare as biased any item which meets any one or more of the four following conditions:

(1) $d(a) > .80$
(2) $d(b) > .50$
(3) $d(c) > .15$
(4) $d(a) + d(b) > 1.00$

The dotted lines in Figures 18.9a, 18.9b, and 18.9c reflect the first three of these criteria. There are so many legitimate objections to these rules of thumb that I shall not try to justify them. I developed them merely by looking at my data and trying to come up with something usable and plausible. Perhaps the outrageousness of my suggestion will motivate the research necessary to develop truly scientific criteria. In the meantime practitioners must practice.

18.12 As a slight digression I feel compelled to mention a study reported by Weiss (1975). The study was only a small part of the cited reference (pp. 33-35), and I shall discuss only a portion of the results. Furthermore, Weiss is very cautious in his interpretion. Nevertheless, the potential implications of the results, if replicable, are of such tremendous import to the field of culture-fair testing that I feel all testing practitioners should be aware of them.

Weiss investigated (among other things) the effect of immediate feedback on test score. He administered a conventional multiple-choice test to Black and White high school students with the items presented by computer on a CRT. (This was not tailored testing. All examinees received the same items.)

Half of each group (Black and White) received immediate feedback from the computer after each response, indicating whether or not the examinee got the item correct. The other half of each group received no feedback.

Feedback was in the form of one of six statements used in a pseudorandom order, such as "right on", "that's cool, now try this one", and "all right, how about this one". The six statements were selected from those suggested by other students at the same high school in order to make the feedback meaningful to the examinees.

With no feedback Blacks scored much worse than Whites, an unfortunate result that has been frequently observed.

However, under the feedback condition Blacks did as well as (actually slightly better than) Whites. Further analysis of the data showed that without feedback Blacks skipped (left unanswered) more items than Whites. But with feedback the Blacks skipped almost no items.

These results suggest that differences in observed test scores between groups may be due to motivational variables, such as a need for encouragement on the part of Blacks, and that, when received, Blacks score as well as Whites.

If these results prove to be replicable, the use of testing machines with appropriate feedback could resolve a large part of the culture-fair testing controversy.

CHAPTER 19
Setting Minimum Passing (Cut-Off) Scores

19.1 One of the more common uses of testing is the classification of examinees into two or more categories. For instance, a college entrance examination may be used to classify into acceptable vs. nonacceptable categories, or remedial program vs. regular program vs. advanced placement categories. A job knowledge test may be used to classify applicants into hire vs. don't hire, or promote vs. don't promote categories. Each of these examples is one of "classification." The examinees are being "classified" into discrete categories.

The classical methodology is to conduct a validation study in which large numbers of persons are tested and measured on the criterion. Then, making the dubious assumption of a linear relationship between test score and criterion measure, the criterion measure is predicted from the test scores. The predictive validity study, the ideal, is almost always extremely expensive and usually impossible in practice. Its less satisfactory alternative, the concurrent validity study, is also usually expensive, and often fraught with problems.

19.2 There are two exciting, inexpensive alternatives to this important and most troublesome psychometric problem, which I shall briefly describe with variations, combining them with IRT. Both techniques are simple to use and rather ingenious.

19.3 Livingston (1976) described a method of finding a criterion-referenced cut-score, which requires only a few criterion measurements The Livingston method follows:

(1) Give the test to a group of examinees.
(2) Pick an examinee with an average test score and measure his performance on the criterion.

(3) If he is competent (satisfactory) on the criterion, pick another examinee with a lower test score. If he is incompetent (unsatisfactory) on the criterion, pick an examinee with a higher score.

(4) Repeat step (3) over and over, each time reducing the difference between the last test score and the next one. (Livingston gives several methods of minimizing the number of criterion measurements required). With each repetition of step (3), the range of uncertainty of the test score that corresponds to the level of minimum competence will be diminished. When you have "zeroed-in" on the minimum test score with sufficient accuracy, you can stop.

Livingston's technique has two significant limitations. First, it uses the number-right score as the predictor. As we have seen in Sec. 12.9, the number-right score can correspond to a wide range of θ, unless the test happens to have high information at the cut-off θ. Since the cut-off θ is not known at the beginning of the technique, finding high information at the ultimately-determined cut-off θ would be pure luck. Selecting examinees on the basis of their θ estimates would improve this method. (This would be an ideal application of tailored testing.) Once the cut-off θ is found, one can redesign the test to have high information at that θ, and then use the corresponding number-right score for selection.

Second, the technique seems to assume that the criterion-measure is unidimensional, and the same dimension as the test. Obtaining a unidimensional criterion measure will be difficult in many practical circumstances.

Both use of the number-right score and nonunidimensionality of the criterion (and/or the test) will result in failure to find a sharp cut-score. Rather, the percent of examinees found satisfactory will rise gradually as test score increases, looking much like an IRF. The decision must then be made of an acceptable risk level of probability of success on the criterion. Usually a 50% risk is used.

In some situations it may be relatively easy to identify initially a group of persons of marginal competency (e.g. "Supervisor, give me a list of your barely acceptable subordinates.") If it is feasible to do so, one may administer the test to them, and find their average θ (or average number-right score, if need be). Their average θ, or score, would be near the best cut-off.

The Livingston technique of selecting examinees with higher or lower θ is analogous to the item selection technique in tailored testing of selecting harder or easier items. Hence, I dub this technique "tailored cutting."

19.4 The other cut-score setting technique, called MAPL,* (Minimum Acceptable Performance Level) was introduced by Nedelsky (1954).

One version of the MAPL procedure follows:

(1) Assemble a group of six to eight subject-matter experts (SMEs).

(2) Instruct the SMEs to form a picture in their minds of the barely acceptable person for the job (or other criterion).

*pronounced "maple"

(3) Each SME then reads each test item and item distractor and asks himself the question, "Would the barely acceptable person know that this distractor (wrong alternative) is wrong?"

(4) If his answer to the question is

(a) definitely no, he assigns two points to the distractor.

(b) definitely yes, he assigns 0 points to the distractor.

(c) neither definitely yes nor no, he assigns one point to the distractor.

(5) Two points are always assigned to the correct choice (key).

(6) Add up the points assigned to all the choices of the item (including the key) by each of the SMEs.

(7) Average the total points assigned to the item by the SMEs.

(8) Divide the average total points into two. The quotient is called the ASI (Alternative Similarity Index).

$$ASI = 2 \div (\text{average total points})$$

(9) Add up the ASIs for all the items in the test. This sum is the MAPL for the test. The MAPL is the number-right score of the minimally acceptable person.

$$MAPL = \sum ASI$$

MAPL is amazingly simple and highly effective in identifying unsatisfactory individuals and/or training need areas. (See Meredith, 1977).

19.5 MAPL may be made even simpler and perhaps more effective by combining it with IRT.

The ASI is, in fact, an estimate of the P(θ) of a person with a barely acceptable θ. This identity may be seen by considering the two extreme cases. If the SMEs assign two points to every distractor in a four-choice item, then the ASI = 2 ÷ (2+2+2+2) = 2 ÷ 8 = 1/4 = .25. The SMEs have, in effect, judged that the barely acceptable person would not know that any of the distractors are wrong, and hence all choices are equally attractive. The barely acceptable person then would have to guess and, assuming he guesses randomly, would have a .25 chance of guessing correctly.

On the other hand, if the SMEs assigned zero points to every distractor, that means the SMEs judged that the barely acceptable person will know all the distractors are wrong. He will thus be sure to get the item correct and the ASI = 2 ÷ (2+0+0+0) = 2 ÷ 2 = 1.00.

If the test is unidimensional and if the items have been pre-calibrated, then, using the ASI as an estimate of P(θ), it is an easy matter to get the θ of the barely acceptable person for that item with the following formula:

$$MAPL\ \theta = b + \frac{1}{1.7a} \log\left[\frac{ASI - c}{1 - ASI}\right]$$

where, log means the natural logarithm. This formula is merely the logistic formula for P(θ), solved for θ, and substituting ASI for P(θ). MAPL θ is the estimated θ of the minimally acceptable person.

In the usual application of MAPL the SMEs must assign points to every distractor of every item in the test. However, in this suggested alternative, the SMEs would only have to do a few items, perhaps 10 to 15. The MAPL θ would then be the average estimated θ of the barely acceptable person for the 10 to 15 items.

This melding of MAPL and IRT also has some limitations. MAPL θ presumes, as does MAPL, that the SMEs are able to properly make the decisions required of them.

Since the ASI cannot be less than .25 (for a four-choice item), it assumes that $c = 1/A$, which we have shown is often not the case. Therefore, it would be well to choose items which have $c \approx .25$.

This technique also has a compounding of error. That is, the ASI is an estimate of $P(\theta)$, and the a, b, and c values are estimates of the true item parameters. When the two estimates are combined, their separate errors may be multiplied. To reduce this compounding of error, items should be chosen which have moderately low a-values, i.e. about $a = 1.00$. Furthermore, the items used should have a range of b-values from, say, -1.00 to +1.00.

MAPL θ is untried, and should be used with caution until adequate research on it can be done. It is, therefore, more of a suggestion than a true alternative.

POSTWORD

The purpose of any communication is the creation of understanding. That is my sole purpose: to create understanding of IRT in the reader.

If there is any part of this publication that you do not understand, then I have not been completely successful in my effort.

Therefore, I would sincerely appreciate any comments, suggestions, questions, corrections, ideas, or discussion about this publication. Please feel free to telephone or write to me for further explanation, discussion, criticism, or just plain chew the fat about IRT.

THOMAS A. WARM, Chief, Exam Branch
Research and Examination Division
U.S. Coast Guard Institute
P.O. Substation 18
Oklahoma City, OK 73169

(405)686-2417 -- commercial
732-2417 -- FTS

APPENDIX A
LOGISTIC IDENTITIES & EQUATIONS

θ = ability parameter
a, b, c = item parameters

e = natural logarithm base

$1.7^2 = 2.89$

Let $x = 1.7a(\theta - b)$

$$e^{-x} = e^{-1.7a(\theta-b)} = \left[\frac{e^{\theta}}{e^{b}}\right]^{-1.7a} = \left[\frac{e^{b}}{e^{\theta}}\right]^{1.7a} = \frac{e^{1.7ab}}{e^{1.7a\theta}}$$

$$\frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

$$\frac{1}{1+e^{x}} = \frac{e^{-x}}{1+e^{-x}}$$

$$\frac{e^{x}}{(1+e^{x})^2} = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$P(\theta) = c + \frac{1-c}{1+e^{-x}} = \frac{c+e^{x}}{1+e^{x}} = \frac{ce^{1.7ab} + e^{1.7a\theta}}{e^{1.7ab} + e^{1.7a\theta}}$$

$$Q(\theta) = 1 - P(\theta) = \frac{1-c}{1+e^{x}}$$

$$P(\theta)\cdot Q(\theta) = \frac{(1-c)\left[c+e^{x}\right]}{\left[1+e^{x}\right]^2}$$

$$P'(\theta) = -Q'(\theta) = \frac{1.7a(1-c)e^{x}}{\left[1+e^{x}\right]^2} = \frac{1.7a(1-c)e^{-x}}{\left[1+e^{-x}\right]^2}$$

$$I(\theta,u) = \frac{P'^2}{P\cdot Q} = \frac{2.89(1-c)a^2}{\left[c+e^{x}\right]\left[1+e^{-x}\right]^2}. \quad \text{If } c = \emptyset,\ I(\theta,u) = 1.7a\,P'(\theta).$$

$$P''(\theta) = \frac{(1.7a)^2(1-c)(1-e^{x})e^{x}}{(1+e^{x})^3}$$

$$W(\theta) = \frac{P'(\theta)}{P(\theta)\cdot Q(\theta)} = \frac{1.7ae^{x}}{c+e^{x}}$$

REFERENCES

Baker, F. B. Advances in item analysis. Review of Educational Research, Winter 1977, vol. 47, No. 1, 151-178

Bejar, I. I. An application of the continuous response level model to personality measurement. Applied Psychological Measurement, vol. 1, No. 4, Fall 1977(a), 509-521.

Bejar, I. I., Wiess, D. J., and Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement. Research report 77-5, Psychometric Methods Program, Dept. of Psychology, Univ. of Minnesota, Sep 1977(b).

Birnbaum, A. "Some latent trait models and their use in inferring an examinee's ability." In Lord, F. M. & Novick, M. R., Statistical Theories of Mental Test Scores; Addison-Wesley, Reading, Mass., 1968; Chpts. 17-20.

Bock, R. D., and Lieberman, M. Fitting a response model for N dochotomously scored items. Psychometrika, 1970, 35, 179-197.

Brogden, H. Variation in Test Validity with Variation in the Distribution of Item Difficulties, Number of Items, and Degree of their INtercorrelation. Psychometrika, 1946, 11, 197-214.

Brown, J. B. and Weiss, D. J. An adaptive testing strategy for achievement test batteries. Research Report 77-6, Oct 1977. Psychometric Methods Program, Dept. of Psychology, Univ. of Minnesota.

Carroll, J.B. Problems in the factor analysis of tests varying difficulty. Amer. Psychologist, 1950, 5, 369. (Abstract).

Cleary, T. A. Test bias: prediction of grades of Negro and White students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Christofferson, A. Factor analysis of dichomotized variables. Psychometrika, 1975, 40, 5-32.

Cronbach, L. J. Essentials of Psychological Testing, Harper and Row, New York, 1960.

Cronbach, L.J. and Warrington, W.G. Efficiency of multiple-choice test as a function of spread of item difficulties. Psychometrika, 1952, 17.

Darlington, R. B. Another look at "cultural fairness." Journal of Educational Measurement, 1973, 10, 237-255.

Durovic, J.J. Use of the Rasch model in assessing item bias. Paper presented a part of a symposeum entitled "What's Happening In Measurement: The Use of Rasch and other Latent Trait Models," Eastern Educational Research Association, Williamsburg, VA, March, 1978.

Ferguson, G. A. Item selection by the constant process. Psychometrika, 1942, 7, 19-29.

Gorman, S. Computerized adaptive testing with a military population. Paper presented at the Computerized Adaptive Testing '77 Conference, Univ. of Minnesota, July 1977.

Green, S. B., Lissitz, R. W., Mulaik, S. A. Limitations of coefficient alpha as an index of test unidimensionality. Education and Psychological Measurement, 1977, 37, 827-838.

Gugel, J. F., Schmidt, F. L., and Urry, V. W. Effectiveness of the ancillary estimation procedure. Proceedings of the First Conference on Computerized Adpative Testing, U.S. Civil Service Commission, Bureau of Policies and Standards. Professional Series, 75-6, 1975, 103-106.

Gugel, J. F. A multiple-choice biserial correlation adjustment for guessing. Unpublished.

Gulliksen, H. Theory of mental tests, John Wiley & Sons, Inc., New York, 1950.

Hambleton, R. K. and Cook, L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hoffman, B. The Tyranny of Testing, Collier, 1962.

Hunter, J. E. and Schmidt, F. L. A critical analysis of the statistical and ethical implications of various definitions of "test bias." Psych. Bulletin, vol. 83, #6, Nov. 1976, 1053-1071.

Kendall, M., and Stuart, A. The Advanced Theory of Statistics, vol. 1, Distribution Theory, MacMillan Publishing Co., New York, 1977

Lawley, D. N. On problems connected with item selection and test construction. Recordings of the Royal Society of Edinburgh, 1943, 61, 273-287.

Livingston, S. A. Choosing minimum passing scores by stochastic approximation techniques. ERIC # ED 135837, Educational Testing Service, Princeton, NJ, Sep 1976.

Lord, F. M. A theory of test scores. Psychometric Monograph, No. 7, 1952

Lord, F. M.. An empirical study of the normality of independence of errors of measurement in test scores. Psychometrika, 1960, 25, 91-104.

Lord, F.M. An empirical study of item-test regression, Psychometrika, 1965, 30, 373-376.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, Sep 1969, vol. 34, #3, 259-299.

Lord, F. M. A study of item bias, using item characteristic curve theory. In Poortingal, Y. H. (ed) Basic Problems in Cross-Cultural Psychology Amsterdam: Swets and Zeitlinger, 1977. 19-30.

Lord, F. M. and Novick, M. R. Statistical Theories of mental test scores. Addison-Wesley, Reading, Mass., 1968.

Lumsden, J. "Test Theory." Annual Review of Psychology. Ed. M. R. Rosenzweig and L. W. Porter. Palo Alto, CA: Annual Reviews, Inc., 1976.

McBride, J. R. and Weiss, D. J. A word knowledge item pool for adaptive ability measurement. Research Report 74-2, Psychometric Methods Program, Dept. of Psychology, Univ. of Minnesota, June 1974.

Meredith, J. B. Jr., and Dion, R. J. Utilization of differential proficiency levels for criterion-referenced training system assessment. Proceedings of a Symposium presented at the 19th Annual Convention of the Military Testing Association, 1977, 1258-1269.

Nedelsky, L. Absolute Grading Standards for Objective Tests. Educational and Psychological Measurement, 1954, 14: 3-19.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Pine, S. M. Applications of item characteristic curve theory to the problem of test bias, in Weiss, D. J. (ed) Proceedings of a Symposium presented at the 18th Annual Convention of the Military Testing Association Research Report 77-1, Psychometric Methods Program, Dept. of Psychology, Univ. of Minn., Mar. 1977, 37-43.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedogogiske Institute, 1960.

Ree, M. J. Implementation of a model adaptive testing system at an AFEES. Paper presented at a symposium of the Eighteenth Annual Conference of the Military Testing Association, San Antonio, Tx., 1977a.

Ree, M. J. Personal communication. 1977b.

Ree, M. J. Automated test item banking. Technical Report AFHRL-TR-78-13, Air Force Human Resources Laboratory, Brooks Air Force Base, Tx, May 1978.

Rudner, L. M. An evaluation of select approaches for biased item identification. Proceedings of a Symposium presented at the 19th Annual Convention of the Military Testing Association, 1977.

Rummel, R. J. Applied factor analysis. Evanston, IL: Northwestern University Press, 1970.

Scheuneman, J. Validating a procedure for assessing bias in test items in the absence of an outside criterion. Paper presented at the American Educational Research Association, San Francisco, April, 1976.

Schmidt, F. L. The Urry method of approximating the item parameters of latent trait theory. Educational and Psychological Measurement, Autumn, 1977, vol. 37, #3, 613-620.

Sympson, J. B. Estimation of latent trait status in adaptive testing procedures. Proceedings of a Symposium presented at the 18th Annual Convention of the Military Testing Association, Research Report 77-1, March 1977, Psychometric Methods Program, Dept. of Psychology, Univ. of Minn., 5-23.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.

Urry, V. W. A Monte Carlo investigation of logistic mental test models. Dissertation Abstracts International, 1971, 31, 6319B.

Urry, V. W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.

Urry, V. W. The effects of guessing on parameters of item discriminatory power. Technical note 75-2, Research Section, Personnel Research and Development Center, U. S. Civil Service Commission, May 1975.

Urry, V. W. Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Weiss, D. J. Adaptive testing research at Minnesota - overview, recent results and future directions. Proceedings of the First Conference on Computerized Adaptive Testing. U. S. Civil Service Commission. Professional Series 75-6, March 1976, pp. 24-25.

Wood, R. L., Wingersky, M. S., and Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.